INVESTIGATING AND MITIGATING THE PERFORMANCE-FAIRNESS TRADEOFF VIA PROTECTED-CATEGORY SAMPLING

by

Gideon Popoola

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Computer Science

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2024

ACKNOWLEDGEMENTS

I would like to thank Dr. John Sheppard, my advisor and committee chair, for his support and guidance throughout this new exciting chapter of my life, and my fellow members of the Numerical Intelligent Systems Laboratory at Montana State University for their valuable comments and suggestions.

I would also like to thank the members of my thesis committee, Dr. Sean Yaw and Dr. Ann Marie Reinhold for all the information and help during the development of this work.

Finally, I cannot begin to express my thanks to my beloved wife, Chioma, for her constant support and inspiration during this thesis writing process, and my family, for their unconditional love and for always being there for me.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF TABLES – CONTINUED

Table Page

LIST OF TABLES – CONTINUED

Table Page

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

Figure Page

LIST OF ALGORITHMS

ABSTRACT

Machine learning algorithms have become common in everyday decision-making, and decision-assistance systems are ubiquitous in our everyday lives. Hence, research on the prevention and mitigation of potential bias and unfairness of the predictions made by these algorithms has been increasing in recent years. Most research on fairness and bias mitigation in machine learning often treats each protected variable separately, but in reality, it is possible for one person to belong to multiple protected categories. Hence, in this thesis, combining a set of protected variables and generating new columns that separate these protected variables into many subcategories was examined. These new subcategories tend to be extremely imbalanced, especially in the class-protected category, so bias mitigation was approached as an imbalanced classification problem. Specifically, four new custom sampling methods were developed and investigated to sample these new subcategories. These new sampling methods are referred to as Protected-Category Oversampling, Protected-Category Proportional Sampling, Protected-Category Synthetic Minority Oversampling Technique (PC-SMOTE), and Protected-Category Adaptive Synthetic Sampling (PC-ADASYN). These sampling methods modify the existing sampling method by focusing their sampling on the new subcategories rather than the class labels. The impacts of these sampling strategies were then evaluated based on classical performance and fairness metrics in classification settings. Classification performance was measured using accuracy, precision, recall, and F1 based on training univariate decision trees, and fairness was measured using equalized odds differences, disparate impact, predictive equality, and statistical parity. To evaluate the impact of fairness versus performance, these measures were evaluated against decision tree depth. The results show that the proposed methods were able to determine optimal points whereby fairness was increased without decreasing performance, thus mitigating any potential performance-fairness tradeoff. To evaluate the impact of the newly proposed sampling on class labels, we also carried out an experiment that calculated the probability distribution of each class of each protected category, and we used KL divergence to measure the difference between these distributions. The results of this experiment show varying KL between the no-sampling and the sampling methods.

CHAPTER ONE

INTRODUCTION

As machine learning (ML) algorithms increasingly dominate decision-making and decision-assistance systems, their widespread deployment across various sectors raises pressing issues about the fairness and transparency of their predictions [64]. The potential for these algorithms to perpetuate or exacerbate existing societal biases has propelled a significant body of research to investigate and mitigate algorithmic unfairness. This is critical because the decisions influenced by these algorithms profoundly impact individuals, affecting outcomes in domains ranging from finance and employment to criminal justice and healthcare [87].

The source of unfairness and bias in ML is multi-faceted [57]. In particular, it is possible that unfairness arises directly from the ML algorithms themselves due to a possible misalignment of the underlying *inductive bias* of the algorithms vis-à-vis the target concept and data distribution. This is referred to as *algorithmic bias*. An alternative concern lies in potential bias resident in the data used to train the models where, as a direct result of the typical "independent and identically distributed" (IID) assumption employed in most ML methods, the result of learning is to propagate the bias in predictions such that they match the bias in the underlying data itself. It is this latter situation that constitutes the focus of our work here.

1.1 Bias and Unfairness in Machine Learning

Bias is the prejudicial, unfair, or unequal treatment of an individual or group based on specific features, often referred to as sensitive or protected features [75]. Examples of these protected features include age, race, disability, sex, and gender [2]. Bias in ML can be divided roughly into disparate treatment (direct unfairness) and impact treatment (indirect unfairness) [53]. Direct unfairness happens when protected features are used explicitly in making decisions to the detriment of corresponding protected groups. Indirect unfairness has become increasingly common today. This type of unfairness does not use protected attributes explicitly; instead, it occurs when reliance on variables correlated with these attributes results in significantly different outcomes for the protected groups. These other variables are known as proxy features. Examples of real-world bias include the historical U.S. practice of "redlining," (where home mortgages were denied to residents of zip codes predominantly inhabited by minorities), Amazon hiring process gender bias, Google soap dispenser racial bias, etc. [18].

Though these decision assistance tools help automate the decision-making process, such tools may result in unfair treatment of either individuals or groups, both directly or indirectly [89]. Unfairness can occur in several areas of modeling, such as in the training dataset. This can happen when the training dataset does not provide a fair representation of the protected categories, so the "ground truth" becomes difficult to determine. For example, consider a dataset from a company where a group has historically faced discrimination. Specifically, suppose female employees in this company have not been promoted at the same rate as their male counterparts, who, in contrast, have seen career advancement despite both groups performing at the same level. In this situation, the true value of female employee contributions—the ground truth—is not visible. As a result, an ML algorithm trained on this data is likely to detect and incorporate this bias, thereby perpetuating existing prejudices.

This could lead to the algorithm making discriminatory decisions, such as recommending male candidates for hire or promotion more frequently than equally or more qualified female candidates.

Another area where unfairness can occur is in the ML algorithm itself [23]. ML algorithms can still produce discriminatory decisions, even when trained on an unbiased dataset where the "ground truth" is represented accurately. This situation arises when the system's errors disproportionately impact individuals from a specific group or minority. For example, consider a breast cancer detection algorithm that exhibits significantly higher false negative rates for black individuals compared to white individuals, meaning it fails to identify breast cancer more frequently in black patients than in white patients. If this algorithm is used to inform treatment recommendations, it would erroneously advise against treatment for a greater number of black individuals than white, leading to racial disparities in healthcare outcomes. This underscores the critical need to ensure that algorithms perform equitably across all groups in terms of their training data and how their errors affect different populations. Results from previous literature have reported several cases of algorithms resulting in unfair treatment, e.g., redlining and racial profiling [74], mortgage discrimination [96], and employment and personnel selection [37].

While considerable efforts have been geared toward addressing bias in ML predictions [44, 59], much of the existing research has focused on mitigating bias for single protected attributes in isolation [100]. For example, on a dataset with two protected attributes, race and sex, most existing approaches can learn either a fair model involving race or a fair model involving sex but not a fair model involving both race and sex [18]. However, real-world identities are not singular; they are complex and multifaceted, with individuals often belonging to multiple protected groups simultaneously [108]. For example, an individual can be discriminated against across several protected attributes such as age, race, and sex simultaneously. This intersectionality can lead to compounded forms of bias and

discrimination that are not addressed adequately by single-variable fairness interventions. Therefore, it is critical to develop methodologies that address personal identities' multi-dimensional nature holistically. This thesis seeks to bridge this gap by considering combinations of protected categories, thereby synthesizing these protected categories into comprehensive multi-category groups, and aims to tackle the layered complexities of bias more effectively using novel protected category sampling methods, thus acknowledging and addressing the multifaceted nature of personal identities and potential biases.

The work presented in this thesis is motivated by the problem of using ML algorithms for decision-making in socially sensitive areas such as auto loan assessment, hiring, or mortgage assessment, working with this situation where an individual can belong to several protected categories. Given a labeled training dataset containing two or more protected features, the method proposed combines these protected attributes and then splits them into new multi-categories. These new categories are likely to be extremely imbalanced and need to be balanced to improve the fairness of the prediction of ML algorithms. Popular sampling methods such as simple over-sampling [109], Synthetic Minority Over-sampling Technique (SMOTE) [22], Adaptive Synthetic Sampling (ADASYN) [41], etc., sample data across class labels, which does not align with the goal of our research of sampling across the new multi-categories. Hence, a new class of modifications of these sampling methods is proposed that can sample across the new category rather than class labels. This new class of modified sampling is called *protected-category sampling*. The resulting proposed protected-category sampling methods are used to sample and balance the new categories before performing classification. The novelty of this work is two-fold. First, the proposed approach combines the protected categories to form new multi-categories that mimic what the identity of a human being looks like in the real world. The second is the modification of existing sampling methods to conform with the sampling of these new categories to make sure that all the new categories have the same number of instances.

For demonstration purposes only, a univariate decision tree was chosen as the classification algorithm. The intent is to demonstrate the effects of the different sampling methods on performance, expecting that similar trends will be exhibited regardless of the underlying learning method. The proposed sampling method was compared to the baseline (unsampled data) using accuracy, precision, recall, and F1 as the classification performance metrics, as well as equalized odds differences, disparate impact, predictive equality, and statistical parity as the fairness metrics. Also, several analyses were performed to show how maximum depth in the decision tree affects both accuracy and fairness and how our new sampling methods affect the class labels

1.2 Research Question

Proceeding from empirical observation that a trade-off sometimes exists between fairness and ML performance [50], this research tries to answer several questions such as,

1. RQ1: Can we the acclaim trade-off between performance and fairness using protected category sampling? The question is trying to tackle the existing empirical research on the trade-off between performance and fairness. Specifically, we want to tackle this by sampling each protected category up to the same samples as the favorable group. We believe combining the protected category and sampling this protected category to have the same number of samples will lower the model bias toward the unfavorable protected groups.

2. RQ2: Can we find a sweet spot in our decision tree where both performance and fairness metrics increase? The goal of this research question is to improve fairness without lowering performance metrics, thus exploring the relationship between level of fit (underfitting through overfitting) and fairness.

3. RQ3: What is the impact of our sampling method on class labels? The goal of this research question is to test whether our sampling method has an impact on class labels or not. Also, we want to quantify the impact of each sampling method on these class labels.

1.3 Hypotheses

In the thesis, we hypothesize the following:

1. H1: We hypothesize that by employing sophisticated protected-category sampling techniques designed for these newly formulated multi-category groups, we can significantly increase model fairness in terms of equalized odds differences without decreasing classification performance in terms of accuracy and F1. This hypothesis tackles the existing trade-off in performance and fairness metrics. It also addresses the multifaceted nature of a human being by combining protected attributes before sampling.

2. H2: We hypothesize that we can significantly increase fairness without decreasing performance by exploring the different depths of the decision trees to identify a sweet spot that favors both metrics.

3. H3: We hypothesize that our new sampling methods can increase the positive labels in the dataset, thereby improving both fairness and performance.

This research challenges existing claims of the existence of trade-offs in fairness and ML prediction. It sets the stage for future explorations into the multi-dimensional nature of identity and discrimination in automated decision systems.

1.4 Contributions

This research focuses on fairness in machine learning, especially in delicate situations like recidivism, loan assessment, and health decisions. The research tackles the existing trade-off by combining the protected categories and using four new sampling methods to sample these new imbalanced protected categories. These new sampling methods were tested on five datasets and using four classification and fairness metrics. The results show that our newly protected category sampling outperforms the baseline(no-sampling), and our sampling methods were also able to improve fairness without any detrimental effect on performance.

1.5 Organization

The remainder of this thesis is organized as follows.

In Chapter 2, we cover the necessary information to make the reader familiar with the various topics discussed in this thesis, including sampling and various types of sampling, the definition of unfairness and fairness metrics, and the decision tree algorithm.

Chapter 3 describes the proposed methodology that was used to tackle bias in this thesis. We also provide details of the five datasets that we used in this thesis, the pre-processing techniques that we applied to our dataset, and the hyperparameter tuning that we used for each algorithm.

In Chapter 4, we present the results of several experiments along with statistical hypothesis tests to test our hypothesis, H1, H2, and H3.

In Chapter 5, the experimental results are discussed, and how each algorithm performs on each dataset and each metric is analyzed. Further results on the impact of tree depth on fairness and accuracy are presented as well. Also, we addressed the question of whether there is an inherent trade-off between fairness and performance. Finally, we discuss the results to check the impact of our sampling methods on the class labels in each protected category.

In chapter 6, we discuss several contributions of this thesis, present the summary, and highlight some limitations of this thesis and how to address them in future works.

CHAPTER TWO

BACKGROUND AND RELATED WORKS

In this chapter, we introduce the main concepts and methods that form the foundation for the work presented in later chapters.

2.1 Sampling

Sampling is a statistical technique or procedure a researcher employs to systematically select a subset of data from a larger population to estimate the characteristics of the whole population [91]. Effective sampling is vital in various fields, including statistics, machine learning, and data science, as it allows researchers and scientists to make inferences or predictions while minimizing the cost and time associated with data collection [101]. In machine learning, sampling techniques are particularly vital when dealing with large datasets or imbalanced class distributions [52].

The mathematical foundation of sampling theory is rooted in probability theory and the law of large numbers. A key concept is the sampling distribution, which describes the probability distribution of a statistic calculated from a random sample. For a simple random sample of size n from a population with mean μ and standard deviation σ, the Central Limit Theorem states that the sampling distribution of the sample mean $\bar{x}$ approximates a normal distribution with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$ as n increases [8].

In practice, various sampling techniques are employed depending on the research goals and population characteristics. Simple random sampling is one of the most straightforward and widely recognized forms [70]. In this method, every member of the population has an equal chance of being selected as part of the sample. This is typically achieved using random number generators or other mechanisms that ensure each individual can be chosen randomly.

Simple random sampling is widely known for its simplicity and fairness, as it minimizes bias and makes the sample representative of the population. The primary advantage of this approach is that the results can be generalized to the entire population. However, it requires gaining access to a list of a larger population, which may not be feasible or unavailable in some scenarios [91].

Stratified sampling divides the population into subgroups (strata) based on specific characteristics, then samples from each stratum [25, 48, 69]. This method ensures that each subgroup is adequately represented within the sample, which can increase the statistical efficiency of the estimation. The population is segmented based on attributes such as age, income, education level, or other relevant criteria, and then simple random samples are drawn from each stratum. Stratified sampling is advantageous when researchers expect variations within particular subgroups to influence the data significantly. This approach enhances precision without increasing cost. The sample size for each stratum is often proportional to its size in the population, given by $nh = (Nh/N) * n$, where nh is the sample size for stratum h, Nh is the population size of stratum h, N is the total population size, and n is the total sample size.

Cluster sampling involves selecting groups (clusters) rather than individuals, which can be more cost-effective for geographically dispersed populations [60]. Cluster sampling is often employed when it is difficult or costly to conduct a simple random sampling of the entire population due to its size or geographical dispersion. Instead of sampling individuals, whole groups or cluster participants are randomly selected. This could involve dividing the population into clusters like schools in a district or neighborhoods in a city and then selecting a few of these clusters at random to provide the sample. While cluster sampling is more economical and more accessible to administer, it typically increases the sampling error and can lead to less accurate results compared to simple random sampling [3]. The design effect in cluster sampling, given by $\text{deff} = 1+\delta(m-1)$, where δ is the intraclass correlation between

the clusters and m is the average cluster size, quantifies the loss in precision compared to simple random sampling.

Systematic sampling is a technique where elements are selected from an ordered sampling frame [67, 103]. The most popular form of systematic sampling is an Nth name selection method where, after the required sample size has been calculated, every Nth record is selected from a list of population members. For instance, we might survey every 10th person on an alphabetical list of conference attendees. This method is exceptionally efficient when a streaming population is involved, such as factory production lines. It is simpler and quicker than simple random sampling but can introduce bias if the list has an underlying pattern that aligns with the Nth selection.

Convenience sampling involves selecting samples based on their availability and ease of access [88, 97]. It is the least rigorous method of sampling, often used in exploratory research where the objective is to get an inexpensive, quick sample that may not be a good representation of the population. For example, a researcher might conduct interviews with people in a shopping mall that happens to be nearby to ask about their opinions on milk consumption. While convenient, this method carries a significant risk of bias, making the results less generalizable to the entire population [32].

Sampling finds applications across numerous fields. In market research, companies use sampling to gauge consumer preferences and market trends without surveying every potential customer [27]. In quality control, statistical process control (SPC) employs sampling to monitor manufacturing processes, using techniques like acceptance sampling plans based on the Acceptable Quality Level (AQL), which determines the maximum number of defects that are acceptable in a product or batch [12]. In environmental science, researchers use spatial sampling techniques like kriging (a statistical interpolation method that estimates values at unknown locations based on known values at nearby locations) to estimate pollutant concentrations across large areas based on measurements at specific points [15]. Political

pollsters rely on carefully designed sampling methods to predict election outcomes, often using techniques like random digit dialing (RDD) for telephone surveys. The accuracy of these applications depends on proper sample design and size, with the margin of error typically calculated as $E = z*\sqrt{(p(1-p)/n)}$, where z is the z-score for the desired confidence level, p is the sample proportion, and n is the sample size.

2.1.1 Oversampling

Oversampling is a technique used to correct imbalances in datasets, particularly in scenarios where certain classes are underrepresented, which can skew the performance of machine learning models [42, 66]. In practice, oversampling increases the number of samples of the minority class in the dataset by replicating existing samples or generating new synthetic samples. This method is especially crucial in many real-world applications where the minority class holds significant importance, such as in medical diagnosis [92], fraud detection [65], spam detection [77], fault diagnosis [107], and natural disaster [31], where failure to accurately predict rare events could have severe consequences. By balancing the dataset, oversampling helps achieve better discrimination and recognition of the minority class, ensuring that the trained model does not unduly favor the majority class. One of the advantages of oversampling is its straightforward implementation and the direct improvement in model accuracy concerning the minority class. It can be particularly effective in preventing overfitting compared to simply collecting more data, which might be costly or infeasible. Figure 2.1 illustrates how oversampling works.

For fairness in machine learning, oversampling can help mitigate bias by improving the model's exposure to underrepresented protected categories, leading to more equitable and just decision-making processes. Oversampling protected attributes for fairness in machine learning is a technique used to address bias and promote equitable outcomes in algorithmic decision-making systems [95]. This approach involves intentionally increasing

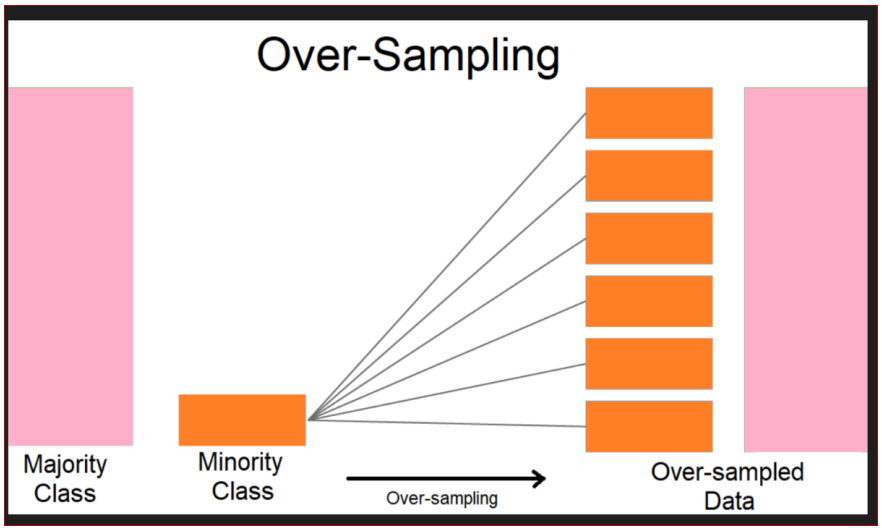

Figure 2.1: Oversampling of Minority Class

the representation of underrepresented or protected groups in the training data to ensure that the model learns to make fair predictions across all demographic categories. The process typically begins with identifying the protected attributes (such as race, gender, or age) and the minority classes within these attributes that are underrepresented in the original dataset. Once identified, additional samples from these protected categories are generated to balance the dataset. This can be done through various methods such as fairness-aware oversampling algorithms based on the distribution of sensitive attributes as proposed in Salazar et al. [86], oversampling for fairness [80], fair oversampling techniques using heterogenous clusters [95], etc.

By oversampling protected attributes using some of the methods described above, the authors were able to expose machine learning models to a more balanced representation of protected categories during training. This can help mitigate bias that might otherwise be learned from imbalanced datasets, leading to more equitable predictions across all groups. However, care must be taken as simple replication oversampling can lead to overfitting, where the model too closely learns the details of the oversampled class. Also, it is important to note

that oversampling is just one tool in the fairness toolkit and should be used in conjunction with other fairness-aware machine learning techniques, careful model evaluation, and ongoing monitoring to ensure truly fair and unbiased algorithmic outcomes.

2.1.2 Porportional Sampling

Proportional sampling is a strategy used in data preparation where the sampling rates for various classes are adjusted to achieve a desired proportional representation that more accurately reflects the target population or meets specific analytical needs [83]. Unlike simple random sampling, which could perpetuate existing class imbalances, proportional sampling ensures that each class in the dataset is represented according to a predefined proportion, thereby aiding in creating a more balanced dataset for model training. This technique is particularly beneficial in scenarios where preserving the relative frequencies of classes is crucial, such as in stratified surveys or population studies where ensuring the representativeness of various subgroups is essential [84]. Figure 2.2 depicts how proportional sampling works. From the figure, the entire population is divided into different subgroups. In each subgroup, proportions are selected according to the number of samples in each subgroup. This helps to lower the influence of groups with larger samples and improve the influence of subgroups with lower samples. In the thesis, we slightly modified this approach by selecting the same number of samples after the protected categories have been divided into different subgroups.

The primary advantage of proportional sampling is its ability to maintain the natural distribution of classes while addressing potential imbalances that could affect the learning process. This method is highly effective in scenarios involving large datasets with significant class disparities, where training a model on the entire dataset is computationally infeasible or could lead to biased outcomes due to overwhelming majority class influence [84]. For instance, in electoral forecasting, proportional sampling can ensure that all demographic

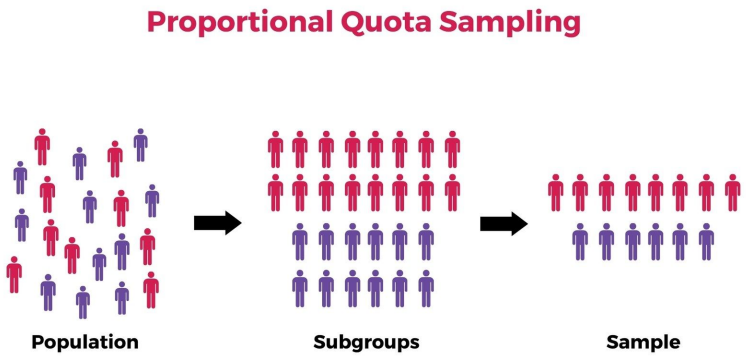

Figure 2.2: Proportional Sampling

groups are adequately represented in the samples, thus providing more accurate predictions of voting outcomes. In the context of fairness in machine learning, proportional sampling can be instrumental in mitigating sampling bias by proportionally representing each protected category, which is crucial for training models that perform equitably across different sensitive attributes [21].

2.1.3 Synthetic Minority Oversampling Technique(SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique developed to tackle the issue of class imbalance in machine learning datasets [11, 22]. Unlike simple replication, which previous oversampling methods are based on, SMOTE generates synthetic samples rather than copying existing ones. SMOTE does this by randomly selecting a point along the line segments that join any of the k minority class nearest neighbors. Depending on the amount of oversampling required, neighbors from the k nearest neighbors are chosen randomly, and synthetic samples are created in the feature space. This method not only helps generate more samples for the minority class but also introduces variability in the samples, thus enabling the model to learn more generalizable patterns rather than memorizing specific instances. This technique is particularly useful in scenarios where the

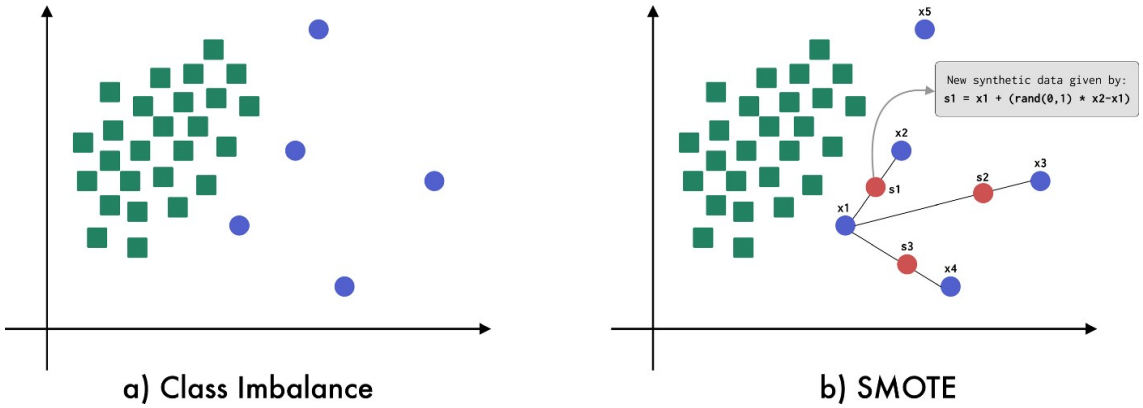

Figure 2.3: SMOTE

minority class is not only underrepresented but also undersampled [6]. Figure 2.3 shows how SMOTE oversampling techniques work.

The advantages of SMOTE are significant, particularly in enhancing the predictive performance of classifiers in imbalanced datasets. By generating synthetic samples, SMOTE allows algorithms to learn more nuanced decision boundaries, which can significantly reduce the bias toward the majority class [61]. This is crucial in fields like fraud detection [68], where fraudulent transactions are rare but must be predicted accurately. In healthcare [79], SMOTE helps in building better predictive models for rare diseases, ensuring that the model is sensitive to subtle cues that might indicate the disease.

When applying SMOTE to protected attributes, care must be taken to ensure that the synthetic samples are meaningful and representative. This may involve adapting the SMOTE algorithm to handle categorical variables that are often present in protected attributes, such as race or gender. Additionally, it is important to validate that the synthetic samples do not introduce new biases or unrealistic data points. Also, SMOTE can sometimes lead to over-generalization, and care must be taken to adjust the number of synthetic samples generated to avoid overfitting.

2.1.4 Adaptive Synthetic Sampling (ADASYN)

ADASYN (Adaptive Synthetic Sampling) is an advanced oversampling technique designed to address the issue of imbalanced data sets in machine learning [41]. Like SMOTE, ADASYN generates synthetic samples for the minority class; however, it focuses more on generating samples next to the borderline than in other regions of the feature space [14]. This core idea of ADASYN is based on the idea that the harder-to-learn examples near the decision boundary are more informative in modifying the decision boundary itself. ADASYN uses a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for those samples that are harder to learn [4]. This results in a more adaptive approach to synthetic sample generation, thereby enhancing the ability of the model to generalize. Figure 2.4 shows how ADASYN oversamples data. In the figure, there are two classes (blue and orange). The blue class is the minority, while the orange class is the majority. The left side image in the figure shows the imbalanced class nature of the dataset. The right side image shows how ADASYN focuses on the samples at the decision boundary and how it tries to replicate the samples at the decision boundary to improve the presence of the minority class and ultimately adjust the decision boundary to its favor, thereby reducing the chance of misclassification of the minority class.

The ADASYN algorithm begins by calculating a density distribution for minority class samples. It determines which minority samples are harder to learn by looking at the ratio of majority to minority samples in their local neighborhood. Samples with a higher ratio of majority neighbors are considered more difficult to learn and are given priority in the synthetic sample generation process. This adaptive approach ensures that more synthetic samples are generated in regions where the minority class is underrepresented and more likely to be misclassified [98].

The advantages of ADASYN are particularly evident in its ability to improve classification accuracy by adaptively changing the learning landscape of the classifier [4]. By

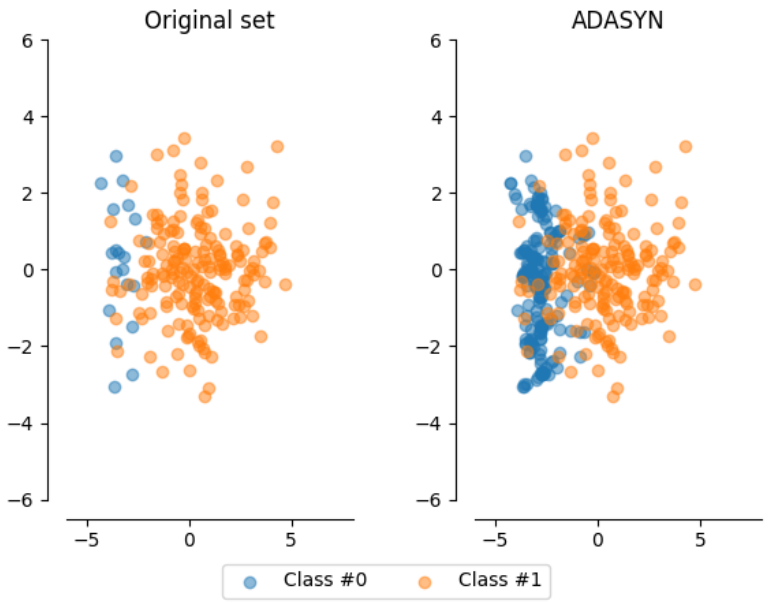

Figure 2.4: ADASYN

focusing on generating more synthetic data around the more difficult instances, ADASYN helps to shift the algorithm decision boundary towards these challenging regions. This method is highly beneficial in applications such as credit risk assessment, where it is crucial to identify potential defaults accurately, and in predictive policing, where it's essential to avoid biased interpretations of demographic data. For fairness in machine learning, ADASYN helps ensure that the predictive models do not neglect the minority class, which often represents underprivileged or less-represented demographics. However, while ADASYN can be highly effective in specific contexts, it can also introduce noise into the dataset if the minority class examples near the boundary are outliers, thus potentially leading to overfitting. It is essential to apply ADASYN in conjunction with robust validation techniques to ensure that the synthetic samples genuinely enhance the model performance and fairness.

When applied to protected attributes in fairness-aware machine learning, ADASYN can help create a more balanced representation of minority groups, particularly in complex

feature spaces where certain subgroups may be especially underrepresented or difficult for the model to learn. By focusing on these challenging areas, ADASYN can potentially improve model performance and fairness across different demographic groups.

2.2 Decision Tree

A decision tree is a popular machine-learning algorithm used for both classification and regression tasks [72]. Decision tree algorithms are built through a process known as recursive partitioning, where data is successively split according to certain conditions. Each node in the tree represents a feature in the dataset, each branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node, which splits into further nodes based on the feature that provides the best homogeneity or separation of the classes in the target variable.

2.2.1 Concepts of Decision Tree

Figure 2.5 shows a simple decision tree model. The decision contains some decision nodes and leaf nodes. Constructing a decision tree involves deciding which features to split on and at what values these splits should occur. This decision is made using mathematical criteria that measure the purity or impurity of the nodes after the split. Examples of popular metrics include Gini Impurity and Entropy. Gini Impurity measures the frequency at which any instance of the dataset will be mislabeled when it is randomly labeled according to the distribution of labels in the subset [99]. Entropy measures the amount of information disorder or the amount of randomness in the data [102]. The decision tree algorithm seeks to partition the data in a manner that decreases the impurity of the nodes and maximizes the information gain, which can otherwise be interpreted as the reduction in entropy or impurity before and after the split.

The choice of where to split the data can be mathematically determined by calculating

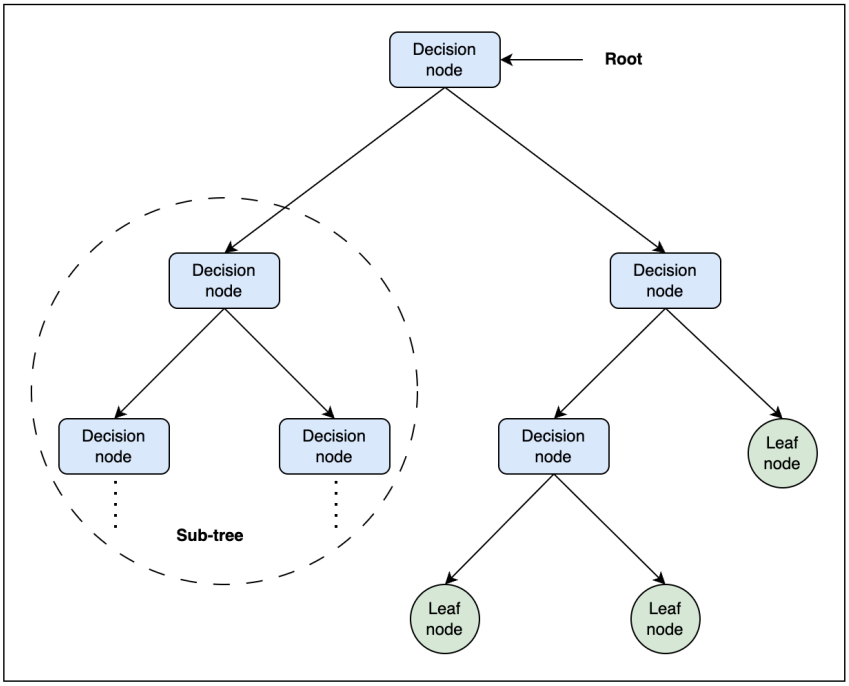

Figure 2.5: Graphical Representation of Decision Tree Algorithm [72]

the Gini impurity or entropy for each possible split across all features and selecting the split that results in the highest gain. For binary classification, Gini impurity can be calculated as

$$Gini = 1 - (p_1^2 + p_2^2 + p_3^2 + ... + p_n^2)$$

Where p_1, p_2, p_3, and p_n are the probabilities of the object being classified to n class, respectively, within the subset. Entropy, on the other hand, is calculated as

$$Entropy = -\sum(p(i) * log2(p(i)))$$

Where pi is the probability of a data point belonging to class i, and i represents a class. The decision rule that results in the highest information gain is chosen as the split rule at each node.

2.2.2 Advantages and Disadvantages of Decision Trees

The decision tree algorithm is a robust algorithm with several advantages, such as they are simple to understand and more interpretable than other non-linear algorithms, and the ability to make a graphical representation of the tree makes the algorithm easy to visualize and communicate the decision-making process [62]. Another advantage of decision trees is that they require little data preparation compared to other algorithms. An example of data preprocessing that decision trees do not require is data normalization, and they can handle both numerical and categorical data. Additionally, decision trees can handle multi-output problems and can indicate which fields are most important for prediction or classification. This inherent feature selection makes decision trees particularly useful in exploratory analysis to understand the factors driving the outcomes [71].

Though the decision tree algorithm possesses several advantages, there exist several disadvantages, such as they are prone to overfitting [13], especially with complex trees that have many layers, because they might model the noise in the training data rather than the intended outputs. This issue can be mitigated by pruning the tree, setting a minimum number of samples required at a leaf node, or setting a maximum tree depth.

2.3 Mathematical Definition of Unfairness

This study considers fairness when predicting an outcome $y \in \mathcal{Y}$ from a set of features $\mathbf{x} \in \boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^d$ and some additional protected attributes $\mathbf{s} \in \boldsymbol{\mathcal{S}} \subseteq \mathbb{R}^p$, such as race, gender, and sex. For example, in loan prediction, $\mathbf{x}$ represents an applicant's financial history, $\mathbf{s}$ is their self-reported race and gender, and y is whether their loan is approved or denied. A prediction model is considered fair if its errors are evenly distributed across protected groups like different races or genders. The class predictions from training data $\mathcal{D}$ are denoted as $\hat{Y}_{\mathcal{D}} := h(\mathbf{x}, \mathbf{s})$ for some $h : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{S}} \to \mathcal{Y}$ from a class in $\mathbf{H}$. The protected attributes $\mathbf{s} \in \boldsymbol{\mathcal{S}}$

in our study are assumed to be binary with a special value n denoting the unprivileged group. For example, $\boldsymbol{\mathcal{S}}$ could be race and n "non-white"; therefore, the binary nature of $\boldsymbol{\mathcal{S}}$ is $\{w, n\}$ where w represents white applicants, who are the privileged group, and n represents non-white applicants, who are the unprivileged group. The definition can be generalized further to non-binary cases. Discrimination in labeled datasets can be defined as given a dataset $\mathcal{D}$, a feature set $\boldsymbol{\mathcal{X}}$, and a protected attribute set $\boldsymbol{\mathcal{S}}$ with domain values $\{w, n\}$. The discrimination in $\mathcal{D}$ with respect to the group $\boldsymbol{\mathcal{S}} = n$ denoted as $dis_{s=n}(\mathcal{D})$ is defined as

$$dis_{s=n}(\mathcal{D}) = \frac{|\{\mathbf{x} \in \mathcal{D} : \mathbf{x}(s) = w, h(\mathbf{x}) = +\}|}{|\{\mathbf{x} \in \mathcal{D} : \mathbf{x}(s) = w\}|} - \frac{|\{\mathbf{x} \in \mathcal{D} : \mathbf{x}(s) = n, h(\mathbf{x}) = +\}|}{|\{\mathbf{x} \in \mathcal{D} : \mathbf{x}(s) = n\}|}$$

The above definition can be translated to the difference in the probability of a subject being in the positive class for each protected attribute domain $\{w, n\}$. Our study extends the above definition by considering datasets $\mathcal{D}$ that contain two or more protected attributes.

2.3.1 Equalized Odds Difference

Four popular fairness metrics are used. The first is *equalized odds difference* (EOD), which measures how discriminative or fair a prediction is. EOD states that a binary classifier $\hat{y}$ is fair if its False Positive Rate (FPR) and True Positive Rate (TPR) are equal across the domain of $\boldsymbol{\mathcal{S}}$ [40]. FPR and TPR with respect to protected attribute $\mathbf{s} \in \boldsymbol{\mathcal{S}}$ with value n can be defined as

$$\begin{aligned} TPR_n(\hat{y}) &= P(\hat{y} = 1 | y = 1, S = n) \\ FPR_n(\hat{y}) &= P(\hat{y} = 1 | y = 0, S = n) \end{aligned}$$

EOD is then defined mathematically as the maximum difference in TPR and FPR across different groups in a protected attribute. That is,

$$EOD = \max_{i \neq j} \{|TPR_i(\hat{y}) - TPR_j(\hat{y})|, |FPR_i(\hat{y}) - FPR_j(\hat{y})|\}.$$

A fair classifier has an EOD of 0, while an unfair classifier has an EOD of 1. Although achieving a fully fair classifier in practice is almost impossible, this research is geared toward improving EOD without decreasing accuracy. Then for EOD,

$$FPR_n(\hat{y}) = P(\hat{y} = 1|y = 0, S = n) = TPR_n(\hat{y}) = P(\hat{y} = 1|y = 1, S = n)$$

and

$$FPR_w(\hat{y}) = P(\hat{y} = 1|y = 0, S = w) = TPR_w(\hat{y}) = P(\hat{y} = 1|y = 0, S = w)$$

To extend the above EOD definition to our multi-category case, the EOD is calculated for each column, and the macro-average of each multi-category is presented as the final EOD.

2.3.2 Statistical Parity

The second metric used to measure fairness in ML prediction is *statistical parity* (SP). SP defines fairness as an equal probability of being classified as positive [54]. This can be interpreted as each group in a protected attribute having the same probability of being classified with a positive outcome.

$$P(\hat{y} = 1|S = w) = P(\hat{y} = 1|S = n).$$

Similar to EOD, this can be extended to multiple classes such that

$$P(\hat{y} = 1|S = i) = P(\hat{y} = 1|S = j), \ \forall i \neq j.$$

From this, SP can be calculated similarly to EOD by considering the pairwise differences across the protected categories.

$$SP = \max_{i \neq j} \{|P(\hat{y} = 1|S = i) - P(\hat{y} = 1|S = j)|\}.$$

2.3.3 Disparate Impact

The third metric used to measure fairness in ML is *Disparate Impact* (DI) [34]. DI measures the ratio of positive outcomes received by the unprivileged group to that of the privileged group. This metric is closely related to statistical parity but focuses on the outcomes of the decision process. A value of 1 indicates perfect fairness, with values less than 1 indicating potential discrimination against the unprivileged group. In contrast to SP, it considers the ratio between unprivileged and privileged groups. Its origins are in legal fairness considerations for selection procedures that sometimes use an 80% rule to define if a process has a disparate impact (ratio smaller than 0.8) or not. Mathematically, DI can be defined as

$$\frac{P(\hat{y} = 1|S = w)}{P(\hat{y} = 1|S = n)}$$

2.3.4 Predictive Equality

The fourth metric is *Predictive Equality (PE)*, which is a fairness metric that asserts that all groups should have equal false positive rates.[9]. This metric can be mathematically expressed as

$$P(\hat{y} = 1|y = 0, S = w) - P(\hat{y} = 1|y = 0, S = n)$$

PE is achieved if this difference is close to zero.

2.4 Related Work

ML algorithms, increasingly utilized for decision-making in critical applications such as recidivism, credit scoring, loan decisions, etc, might initially be assumed to be fair and free of inherent bias. However, in reality, they may inherit any bias or discrimination present in the data on which they are trained, as noted by Burt [17]. Moreover, merely removing protected variables from the dataset is insufficient to tackle indirect discrimination and might, in fact, conceal it. This recognition has heightened the need for more advanced tools, making discovering and preventing discrimination a significant area of research, as highlighted by [16, 93].

Bias in ML is a fast-growing topic in the machine learning research community. Bias in an ML model can lead to an unfair treatment of people belonging to certain protected groups. Lately, industrial leaders have started putting more emphasis on bias in ML models and software. The Institute for Electrical and Electronics Engineers (IEEE) [90], Microsoft [24], and the European Union [30] have recently published principles for guiding fair AI conduct. These organizations have stated that ML models must be fair in real-world applications. Bias mitigation strategies involve modifying one or more of the following to ensure the predictions made by the ML algorithm are less biased: (a) the training data, (b) the ML algorithm, and (c) the ensuing predictions themselves. These are respectively categorized as pre-processing [39], in-processing [35], and post-processing approaches [40].

2.4.1 Preprocessing

First, the training data can be preprocessed to lower unfairness or bias before training the model. Kamiran and Calders suggest sampling or re-weighting the data to neutralize discrimination [53]. This approach can adjust the representation or importance of certain data points to favor (or reduce favor) one class over another. Another method involves

changing individual data records directly to reduce discrimination, as explored by [19]. For example, this approach involves altering values in a dataset to decrease identifiable biases against certain groups. Additionally, the concept of t-closeness, a privacy model that reduces the granularity of a data representation to protect against attribute disclosure, introduced by Sondeck *et al.* [94], is applied to discrimination control in the work of [85]. Using t-closeness ensures that the distribution of sensitive attributes in any given group is close to the distribution of the attribute in the entire dataset, thereby preserving privacy and preventing discrimination based on sensitive attributes. A common thread among these approaches is balancing discrimination control with the processed data's utility, that is, minimizing bias without significantly compromising the data's accuracy, representativeness, and overall usefulness for predictive modeling or analysis. This balance is essential for ensuring that efforts to promote fairness do not inadvertently reduce the quality or applicability of the data.

Overall, the pre-processing method can be divided further into three categories: 1) data modification, 2) data removal, and 3) data resampling. Methods in the first category aim to modify the values of the training data points (including protected attribute values, class values, and feature values) to lower the bias in the dataset. An example of this method is data massaging proposed by [33]. Their approach ranks the training data, and data close to the decision boundary in both privileged and unprivileged groups are interchanged. Alternatively, an optimized pre-processing method that learns a probabilistic transformation that edits the classes and features with individual distortion and group fairness was proposed by Fahse *et al.* [17]. In [82], the original attribute values are replaced with values chosen independently from the class label to train a model roughly achieving equalized odds. Similarly, Peng *et al.* replace the protected attribute values with values predicted based on other attributes, similar to data imputation [73].

Methods in the second category aim to train a fair model by removing certain features

from the training set. An example of this method is data suppression proposed by Dhar *et al.* [26]. In their paper, the protected attributes and features that are highly correlated with protected attributes, otherwise known as proxy attributes, are removed from the dataset to train a fair model.

Methods in the third category aim to train a fair model either by adjusting the sample weights or by oversampling the dataset. For example, Krasanakis *et al.* proposed a reweighting method that iteratively adapts training sample weights with a theoretically grounded model to mitigate the bias-accuracy tradeoff [58]. In [81], Chakraborty *et al.* proposed FairSMOTE as a method to over-sample training points from minority groups with artificial data points based on SMOTE [22] to achieve balanced class distributions. Also, Yan *et al.* proposed over-sampling the training data from the minority groups with artificial data points to achieve balanced class distributions [104]. Unlike FairSMOTE, the authors focused on scenarios where protected attributes are unknown and applied a clustering method to identify different demographic groups.

2.4.2 In-Processing

In-processing involves methods that modify the way an ML model is trained as a means to reduce bias. In [47], an adversarial debiasing approach was proposed. This approach learns a classifier to increase accuracy and fairness in prediction by including a variable for the group of interest and simultaneously learning a predictor and an adversary. This leads to the generation of an unbiased classifier because the predictions do not contain any group discrimination information that the adversary could utilize. Alternatively, an algorithm that takes a fairness metric as part of the loss function and returns a model trained for that fairness metric was proposed in [20]. Kamishima *et al.* proposed a regularization method, which included a penalty term in the loss function of a classifier to produce an unbiased prediction [54]. Zafar *et al.* developed a new weighting method whereby they tune

the sample weight for each training datum to achieve a specific fairness objective, such as equalized odds on the validation data [106]. Recently, bias mitigation has been approached as a constrained optimization problem by adding a fairness constraint and optimizing the loss to be consistent with that constraint [1, 63].

2.4.3 Post-Processing

Post-processing methods mitigate bias after fitting an ML model and include approaches such as calibration, constraint optimization, transformation, and thresholding [53]. Such methods yield algorithms that give favorable outcomes to unprivileged groups and less favorable outcomes to favorable groups within a given confidence interval around the decision boundary with the highest uncertainty. For example, one approach modifies the peak thresholds of the classifier to yield a specified statistical parity or equalized odds target. Yet another approach involves randomly mutating the classes of certain predictions into different classes [45].

Several new studies [10, 49] combined either pre-processing, in-processing, or post-processing to form a hybrid method. For example, Bhaskaruni *et al.* combine oversampling the imbalance protected class with a decision boundary shifting a post-processing method to tackle the unfairness problem [49]. Researchers have delved into various concepts of discrimination and fairness within algorithmic decision-making. For example, DI is measured through statistical parity and group fairness, as discussed by Bhaskaruni *et al.* [49]. On the other hand, the concept of individual fairness, also introduced by Bhaskaruni *et al.*, emphasizes that similar individuals should be treated similarly, regardless of their group affiliation. This approach focuses on fairness at the individual level, ensuring that decisions are made based on relevant attributes rather than group-based stereotypes or biases.

In classifiers and other predictive models, achieving equal error rates across different groups is a key goal, as highlighted by Zhang and Neill [108]. Similarly, ensuring calibration

or the absence of predictive bias in the predictions, as discussed by Hardt *et al.*, is the goal [40]. However, the tension between these notions—calibration and equal error rates—is explored by Dwork *et al.*[29] and Pleiss *et al.* [76], indicating that simultaneously satisfying both can be challenging. Karimi-Haghighi and Castillo present related work exploring the complexities inherent in achieving algorithmic fairness [55]. Friedler *et al.* further examines the trade-offs in meeting various algorithmic fairness definitions, especially from a public safety perspective [36]. Given that our work focuses on pre-processing rather than modeling, considerations such as balanced error rates and predictive bias become less directly applicable.

Based on our review of various pre-processing methods, it appears that no work has been done attempting to model fairness for two or more protected attributes simultaneously. Also, the sampling methods used in prior work focused only on sampling based on class labels rather than the protected categories. Hence, in this paper, pre-processing is emphasized as it represents the most adaptable aspect of the data science pipeline [38]. Pre-processing is distinct in that it does not depend on the choice of modeling algorithm and can be incorporated seamlessly with data release and publishing mechanisms. This independence and flexibility make pre-processing critical for ensuring data quality and fairness before any analytical or predictive modeling occurs. Finally, we focus on new custom sampling methods that sample the protected category in the data training to build a fair model.

CHAPTER THREE

METHODOLOGY AND EXPERIMENTAL DESIGN

In this chapter, we describe the proposed algorithms that we will use in Chapter 4 to address unfairness in machine learning predictions. We will also describe the dataset, dataset selection process, hyperparameter tuning, and experimental design used to mitigate the trade-off between bias and fairness.

3.1 Proposed Algorithms

The focus of our work is to explore sampling methods to enhance fairness in ML without the corresponding prediction performance suffering, thus mitigating the fairness-performance tradeoff. As a result, four novel sampling methods focused on achieving this goal are proposed. These sampling methods address the imbalanced class problem posed by the new multi-category generated from the combination of the protected categories. The new multi-categories are generated when we combine two or more protected attributes together, and we perform one-hot encoding to split them into multi-category protected attributes. Custom sampling methods are needed because the existing methods sample data based on minority and majority classes, but to mitigate fairness, the new multi-categories are sampled to be equal. This, in turn, calls for modifying the existing sampling methods to sample data based on these new categories. This leads to four new sampling methods: protected-category over-sampling, protected-category proportional sampling, protected-category SMOTE (PC-SMOTE), and protected-category ADASYN (PC-ADASYN).

3.1.1 Protected-Category Over-Sampling

In Protected-Category Over-sampling, the first step is to combine the protected categories in the dataset and encode the combination to produce our new multi-category.

For example, in the Adult Income dataset, age (young and adult), race (white and others), and sex (male and female) are combined to generate eight new categories, which become AdultWhiteMale, AdultOthersMale, YoungWhiteMale, YoungOthersMail, AdultWhiteFemale, AdultOthersFemale, YoungWhiteFemale and YoungOthersFemale respectively. These new categories have varying sample sizes, and the goal of our protected-category over-sampling is to balance these new categories such that the sample size of each of the new categories matches the size of the category with the largest sample size. To avoid data leakage in our experiments, the dataset is separated into train and test using 10-fold stratified cross-validation, applying over-sampling only on the training data and then testing on an unsampled test set.

The pseudocode in Algorithm 3.1 shows our protected category over-sampling method in detail. In the algorithm, the largest category was used as the baseline because it has the highest sample size. The sampling process results in new training data with a balanced sample size across the new multi-category protected attributes. From line 1 to line 5, the algorithm tries to calculate the multi-category with the highest number of samples, and this category is selected as the baseline. From line 6 to line 13, the algorithm works by sampling the rest of the protected categories (with replacement) to match the sample size of the baseline. This sampling is done by repeating the categories multiple times along with their class labels. Lines 15 and 16 append the new multi-category sample to the original data and return a new sampled dataset with an equal number of samples across the multi-categories.

3.1.2 Protected-Category Proportional Sampling

The Protected-Category Proportional Sampling method is a generalization of protected-category over-sampling because the process begins by setting a target sample size (which is a hyperparameter to be tuned, rather than just the size of the largest multi-category), denoted as $targetSamples$. This corresponds to the desired number of instances needed for

Algorithm 3.1 Protected-Category Over-Sampling

1: **procedure**
2: $baselineCount \leftarrow$ sum of entries in 'Largest_Category' of X_{train}
3: $totalCount \leftarrow$ number of entries in X_{train}
4: $baselineProportion \leftarrow baselineCount/totalCount$
5: $balancedData \leftarrow$ initialize an empty dataset
6: $categories \leftarrow$ list of column names in X_{train} starting with 'combined_category'
7: **for** each $category$ in $categories$ **do**
8: $categoryData \leftarrow$ select entries in X_{train} where $category = 1$
9: $categoryData \leftarrow$ combine $categoryData$ with corresponding labels from y_{train}
10: $categoryCount \leftarrow$ number of entries in $categoryData$
11: $targetCount \leftarrow$ integer part of $totalCount \times baselineProportion$
12: **if** $categoryCount < targetCount$ **then**
13: $sampledData \leftarrow$ sample $targetCount$ from $categoryData$ with replacement
14: $balancedData \leftarrow$ append $sampledData$ to $balancedData$
15: **else**
16: $balancedData \leftarrow$ append $categoryData$ to $balancedData$
17: **return** $balancedData$

each category. This target ensures uniformity across all categories, mitigating the risk of model bias towards more frequent categories. The typical result of applying this method is that some categories that have more samples than the *targetSamples* will be under-sampled while others will be over-sampled to yield an equal proportion of them in the training dataset. The pseudocode in Algorithm 3.2 shows the steps of the protected-category proportional sampling method. Lines 1-5 select the target sample and initialize the empty array for the new sampled dataset. Lines 6-8 sampled the dataset to match the target samples. Lines

Algorithm 3.2 Protected-Category Proportional Sampling

1: **procedure**
2: $targetSamples \leftarrow n$
3: $sampledBalanced \leftarrow$ initialize an empty data set
4: **for** each *column* in *new_categories*.columns **do**
5: $categoryRows \leftarrow$ select rows in *new_categories* where $column = 1$
6: $sampledRows \leftarrow$ sample $targetSamples$ entries from $categoryRows$ with replacement
7: **for** each *col* in *oneHotEncodedBalanced*.columns **do**
8: $sampledRows[col] \leftarrow 0$
9: $sampledRows[column] \leftarrow 1$
10: $sampledBalanced \leftarrow$ append $sampledRows$ to $sampledBalanced$
11: **return** $sampledBalanced$

9 and 10 append the sampled rows to the original dataset and generate a new balanced dataset.

3.1.3 Protected-Category SMOTE

The Protected-Category Synthetic Minority Oversampling Technique (PC-SMOTE) method is a more complex process aimed at mimicking SMOTE but modified for sampling over the new categories rather than the class labels. In this approach, the first step is to modify SMOTE to use a fixed number of neighbors and to randomly select one neighbor for the interpolation rather than averaging all of them. The pseudocode in Algorithm 3.3 shows the procedure for PC-SMOTE. Lines 1-4 create an empty array and fit nearest neighbor on the dataset. Lines 5 to 13 randomly select one neighbor and perform interpolation; then, line 9 adds a random number between 0 and 1 to the interpolation to generate new samples

Algorithm 3.3 Custom Synthetic Minority Over-sampling Technique (SMOTE)

1: **procedure** CUSTOMSMOTE(*data*, *n_samples*)
2: $syntheticSamples \leftarrow$ zero matrix of size ($n\_samples\times$ number of columns in *data*)
3: $nn \leftarrow$ NearestNeighbors($n\_neighbors = n$).fit(*data*)
4: $neighbors \leftarrow nn$.kneighbors($data$, return_distance $= False$)
5: **for** $i \leftarrow 1$ **to** $n\_samples$ **do**
6: $sampleIdx \leftarrow$ random integer from 0 to (number of rows in $data - 1$)
7: $nnIdx \leftarrow$ random choice from $neighbors[sampleIdx, 1:]$
8: $diff \leftarrow data[nnIdx] - data[sampleIdx]$
9: $weight \leftarrow$ random number from uniform distribution between 0 and 1
10: $syntheticSamples[i] \leftarrow data[sampleIdx] + weight \times diff$
11: **return** *syntheticSamples*

in line 10. Since the method's intent is to be used for new category sampling, it does not directly address the generation of class labels. Hence, a new function that can generate a new class label for the synthetic data is needed. For this, a new function is defined that generates class labels based on the number of new synthetic data generated and a preselected balance ratio between the two classes. Algorithm 3.4 shows how our new function generates labels for our synthetic samples. In lines 2 and 3, the algorithm first determines the number of samples for each class based on the balance ratio and generates the sample needed for each class. Line 4 generates this synthetic label, while in line 5, the class labels are then shuffled to prevent algorithmic bias in the classes generated.

These two algorithms are combined together to form PC-SMOTE, as shown in Algorithm 3.5. In lines 3 to 5, the algorithm iterated the dataset over such that the subset of data associated with that category is identified. Line 6 calculates the number of synthetic samples needed to reach a predefined maximum size per category. If additional samples are

Algorithm 3.4 Generate Balanced Synthetic Labels

1: **procedure** GenBalSynthLabels($n\_samplesNeeded$, $balanceRatio$)
2: $nClass1 \leftarrow \text{int}(n\_samplesNeeded \times balanceRatio)$
3: $nClass0 \leftarrow n\_samplesNeeded - nClass1$
4: $syntheticLabels \leftarrow [class0] \times nClass0 + [class1] \times nClass1$
5: Shuffle($syntheticLabels$)
6: **return** $syntheticLabels$

required, in lines 7 to 14, the data is generated using PC-SMOTE, which interpolates between existing data points and their nearest neighbors. Concurrently, a balanced distribution of synthetic class labels is created with a specified balance ratio by employing Algorithm 3.4. These synthetic features and labels are then incorporated into the training subset for each category in line 15. The process is repeated for all categories, resulting in a balanced dataset.

3.1.4 Protected-Category ADASYN

The Protected-Category Adaptive Synthetic Sampling (PC-ADASYN) method mimics Adaptive Synthetic Sampling (ADASYN) sampling but is modified slightly to fulfill our goal of protected category sampling. Our PC-ADASYN algorithm is shown in Algorithm 3.6. It extends ADASYN by focusing on category density rather than class imbalance. Specifically, line 2 to 9 of this function operates by finding the nearest neighbors to the data and then calculating the density of each data point's category within its immediate neighborhood. In line 10-13, it assign weights as inverse density of each point to prioritize minority categories, making it more likely to generate synthetic samples from underrepresented categories. In lines 14 to 18 the synthetic samples are created by interpolating between selected data points and their neighbors, similar to SMOTE but using a random weight to vary the interpolation, thus ensuring a diverse synthetic dataset. This approach helps address the imbalance at the

Algorithm 3.5 Protected-Category SMOTE

1: Procedure
2: $balancedDataList \leftarrow$ initialize an empty list
3: **for** each *category* in *categories* **do**
4: $categorySubset \leftarrow$ select rows in *train_data* s.t. 'combined_category' == *category*
5: $features \leftarrow$ remove 'class', 'combined_category' from *categorySubset*
6: $nSamples \leftarrow max\_size-$ number of rows in *categorySubset*
7: **if** $nSamples > 0$ **then**
8: $syntheticFeatures \leftarrow$ CustomSMOTE($features, nSamples$)
9: $syntheticLabels \leftarrow$ GenBalSynthLabels($nSamples, balanceRatio$)
10: $syntheticFeatures['class'] \leftarrow syntheticLabels$
11: $syntheticFeatures['combined\_category'] \leftarrow category$
12: $categorySubsetBalanced \leftarrow$ concatenate *categorySubset* and $syntheticFeatures$
13: **else**
14: $categorySubsetBalanced \leftarrow categorySubset$
15: append $categorySubsetBalanced$ to $balancedDataList$
16: $balancedData \leftarrow$ append $balancedDataList$ and reset index

category level and enriches the dataset's variance, potentially improving the robustness and fairness of ML models trained on this data. Since this sampling method also generates new samples by interpolating, Algorithm 3.4 is used to generate class labels for the new synthetic samples.

Algorithm 3.4 and Algorithm 3.6 are combined to form PC-ADASYN sampling as shown in Algorithm 3.7. In lines 1 to 5, in each category, the data corresponding to that category is isolated, and the size deficit relative to the largest category is computed. If additional samples are needed, lines 7 to 16 of the PC-ADASYN method are applied, generating synthetic data

Algorithm 3.6 Custom-ADASYN for Category-Based Balancing

```
1: procedure PCAdasynCategories(data, labels, n_samplesNeeded, n_neighbors)
2:     n_neighbors ← n_neighbors + 1
3:     nn ← NearestNeighbors(n_neighbors).fit(data)
4:     distances, indices ← nn.kneighbors(data)
5:     densities ← zero array of length(data)
6:     for i ← 0 to length(data) − 1 do
7:         current_category ← labels[i]
8:         neighbor_indices ← indices[i − 1]
9:         densities[i] ← Sum(labels[neighbor_indices] == current_category)
10:    weights ← 1/(densities + 1)
11:    weights ← weights/Sum(weights)
12:    syntheticSamples ← empty list
13:    sampleIndices ← random choice with replacement from length(data) using weights
14:    for each idx in sampleIndices do
15:        baseIdx ← idx
16:        neighborIdx ← RandomChoice(indices[baseIdx])
17:        diff ← data[neighborIdx] − data[baseIdx]
18:        syntheticSample ← data[baseIdx] + Random() × diff
19:        append syntheticSample to syntheticSamples
20:    return array(syntheticSamples)
```

that respect the category's distribution characteristics. These data are then completed with synthetically generated labels, maintaining a predefined class balance ratio. The process not only corrects category imbalances but also enriches the dataset, potentially enhancing the predictive accuracy and fairness of models trained on these data.

Algorithm 3.7 Protected-Category ADASYN

procedure
    $balancedDataList \leftarrow$ initialize an empty list
    for each *category* in *categories* **do**
        $categorySubset \leftarrow$ select from *data* s.t. 'combined_category' == *category*
        $features \leftarrow$ remove 'class', 'combined_category' from $categorySubset$
        $categoryLabels \leftarrow$ extract 'combined_category' from $categorySubset$
        $nSamplesNeeded \leftarrow max\_size$ minus number of rows in $categorySubset$
        if $nSamplesNeeded > 0$ **then**
            $syntheticFeatures \leftarrow$ Custom-ADASYN($features, categoryLabels, nSamplesNeeded$)
            $syntheticLabels \leftarrow$ GenBalSynthLabels($nSamplesNeeded, balanceRatio$)
            $syntheticFeatures['class'] \leftarrow syntheticLabels$
            $syntheticFeatures['combined\_category'] \leftarrow category$
            $categorySubsetBalanced \leftarrow$ concatenate $categorySubset$ with $syntheticFeatures$
        else
            $categorySubsetBalanced \leftarrow categorySubset$
        append $categorySubsetBalanced$ to $balancedDataList$
    $balancedData \leftarrow$ concatenate $balancedDataList$ and reset index

3.2 Experimental Design

To test the four sampling methods, a classifier was needed to assess the effects on fairness and performance. Ultimately, the type of classifier is not directly relevant since the goal is to mitigate the fairness-performance tradeoff rather than to find the best classifier. Therefore, we chose to use univariate decision trees based on CART [62] due to their robustness against

noise and missing data. In addition, decision trees allow us to control the strength of fit by setting the tree depth of the learned tree. This allows us to compare fairness and performance across different levels of fitting.

The process begins by combining protected categories within each dataset, applying one-hot encoding to create new multi-category features, and then performing label encoding. The datasets are then divided using a stratified 10-fold cross-validation to ensure a representative distribution of classes in each fold. For each fold, training is conducted on sampled data using the previously described methods, while classification is tested on the corresponding unsampled test sets. Consistency in model training is maintained by applying identical tree depth across all sampling methods, and the results provided are averages of the 10-fold runs with their corresponding confidence intervals.

3.2.1 Dataset Selection Criteria

In order to test the efficacy of our new sampling methods, we rigorously selected five datasets using the following criteria.

1. These datasets were selected because they represent state-of-the-art datasets for measuring bias and discrimination and are widely used in other studies on algorithmic bias and fairness.

2. These datasets were selected to include the most common domains of life where unfairness can have the most detrimental effect. The domains that our dataset represents include finance, health, and crime.

3. The datasets have various sizes ranging from small to large, which makes them suitable for testing our sampling methods across different conditions.

4. These datasets are selected from reputable sources (UCI and Kaggle) to ensure reproducibility in our experiments.

3.2.2 Dataset Description

Five datasets were selected from the UCI repository [56] and Kaggle for our analysis: the Adult Income dataset [7], the German Credit [43] dataset, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset [28], the credit card default dataset [105], and CDC Diabetes Health Indicators dataset [51]. The Adult Income dataset aims to predict whether an individual earns above $50,000, featuring eight categorical and four numerical attributes, with protected variables corresponding to age (young or adult), sex (male or female), and race (white or others). The German Credit dataset, used to predict creditworthiness, comprises 20 categorical and two numerical attributes, with protected variables of age and sex. The COMPAS dataset, which assesses recidivism rates in the United States, includes six categorical and six numerical features, with protected variables of age, race, and sex. The credit card default dataset aims to predict whether some Taiwanese customer will default on their credit card. The protected categories in the dataset are sex and age. The CDC diabetes dataset aims to predict whether someone is either diabetic or not. The dataset contains 21 features, of which 18 are categorical and three are numerical. The protected attributes in the dataset are age and sex.

3.2.3 Hyperparamter Tuning

Hyperparameter tuning was conducted using grid search [46] to explore a broad range of parameters, complemented by visual assessments to identify optimal settings that balance Equalized Odds Difference (EOD) and accuracy. For the Adult Income dataset, the optimal hyperparameters included a maximum tree depth of 3 and, for PC-SMOTE and PC-ADASYN, a nearest neighbor setting of 5 with a balanced ratio of 0.34. These parameters were similarly effective for the German Credit dataset and the credit card default dataset. For the CDC diabetes dataset, the hyperparameter includes a maximum tree depth of 7. For the COMPAS dataset, a maximum tree depth of 2 was optimal for all sampling methods.

PC-SMOTE and PC-ADASYN were adjusted to the nearest neighbor setting of 3 and a balanced ratio of 0.60.

CHAPTER FOUR

RESULTS

The four sampling strategies were applied to the five datasets described in chapter 3, and their impact was evaluated using a simple univariate decision tree classifier. The results in Tables 4.1, 4.2, 4.3, 4.4, and 4.5 show notable differences in model performance on the five datasets across five sampling strategies: no sampling, over-sampling, proportional sampling, PC-SMOTE, and PC-ADASYN. Each method was assessed based on accuracy, macro precision, macro recall, macro F1, Equalized Odds Difference (EOD), Statistical Parity (SP), Disparate Impact (DI), and Predictive Equality (PE).

To ascertain the statistical significance of each method's results, we used the Friedman test, a non-parametric alternative to the one-way ANOVA with repeated measures. Upon finding significant results from the Friedman test, we proceeded with the Nemenyi post-hoc test. This test is used to evaluate pairwise comparisons between the methods to ascertain which methods statistically differ from each other. The Nemenyi test is advantageous in this setting because it accounts for multiple comparisons without assuming normal distributions, thereby providing a robust way to understand specific pairwise differences.

4.1 Adult Income Results

For the Adult Income dataset results in Table 4.1, the no-sampling method yielded an accuracy of 0.82, setting a high baseline for comparison. The no-sampling method also yielded a macro-precision of 0.84, a macro-recall of 0.64, and a macro-F1 of 0.64. However, it demonstrated a slightly biased prediction with an EOD of 0.36 and a minimal disparity of 0.02 in SP. Additionally, the DI was very low, at 0.00, indicating severe disparity in positive outcome rates between protected and unprotected groups. The PE was also quite

Table 4.1: Results of the sampling methods on the Adult Income dataset with 95% confidence intervals. **Bolded** results indicate statistical significance.

Metric	No Sampling	Over-sample	Prop. Sample	PC-SMOTE	PC-ADASYN
Accuracy	0.82 ± 0.00	0.82 ± 0.02	0.79 ± 0.05	0.81 ± 0.05	0.82 ± 0.04
Precision	0.88 ± 0.02	0.88 ± 0.02	0.80 ± 0.02	0.88 ± 0.03	0.87 ± 0.02
Recall	0.64 ± 0.03	0.63 ± 0.03	**0.80** ± 0.02	0.63 ± 0.03	0.63 ± 0.03
Macro F1	0.66 ± 0.01	0.65 ± 0.06	**0.79** ± 0.06	0.65 ± 0.07	0.64 ± 0.07
EOD	0.36 ± 0.18	0.66 ± 0.20	0.46 ± 0.12	**0.25** ± 0.21	0.28 ± 0.19
SP	**0.02** ± 0.00	0.09 ± 0.02	0.71 ± 0.10	0.07 ± 0.05	0.09 ± 0.02
DI	0.00 ± 0.03	0.00 ± 0.04	**0.07 ± 0.10**	0.00 ± 0.03	0.00 ± 0.05
PE	0.01 ± 0.03	**0.01 ± 0.02**	0.46 ± 0.15	0.01 ± 0.03	0.01 ± 0.04

low, at 0.01, reflecting the minimal disparity in false positive rates, but this could suggest an overall bias in how positive outcomes are distributed across groups. The discrepancy between these fairness metrics shows how hard it is to optimize for multiple fairness metrics simultaneously. Also, these fairness metrics have different definitions, which means some contradict each other.

In contrast, over-sampling maintained the same accuracy and precision but lowered the macro F1 slightly to 0.65 and recall to 0.64, indicating potential overfitting issues while worsening fairness with an EOD of 0.66 and increasing SP to 0.09. DI remained at 0.00, signaling a severe imbalance in the ratio of positive outcomes between groups, while PE stayed at 0.01, showing minimal change in false positive disparities.

Proportional sampling decreased accuracy to 0.79 and precision to 0.80 but improved the recall to 0.80 and macro F1 to 0.79, indicating a better balance between precision and recall.

However, it significantly increased SP to 0.71, indicating a substantial disparity in positive prediction difference across each group, which raises concerns about the model's fairness. Interestingly, DI improved to 0.07, reflecting a small shift toward a fairer distribution of positive outcomes between groups, while PE increased to 0.46, indicating a higher disparity in false positive rates across groups.

PC-SMOTE and PC-ADASYN were designed specifically to improve upon these metrics. PC-SMOTE showed a comparable performance with an accuracy of 0.81, precision of 0.88, and recall of 0.63. The method yielded an improved EOD of 0.25, suggesting enhanced fairness over basic over-sampling and the no-sampling method. However, it still recorded a lower macro F1 of 0.65, indicating misrepresentation issues in synthetic data generation and a possible trade-off between accuracy and recall. DI remained at 0.00, showing no improvement in the balance of positive outcomes and reflecting bias in the model's prediction, and PE stayed low at 0.01, reflecting minimal disparity in false positives.

PC-ADASYN proved to be the most balanced approach, maintaining a high accuracy of 0.82, a slightly lower precision of 0.87, a recall of 0.63, and a macro f1 of 0.64. It moderately improved fairness EOD = 0.28 and controlled the increase in SP to 0.09. Like PC-SMOTE, DI stayed at 0.00, highlighting the persistent challenge of balancing positive outcome distributions. PE was similarly low at 0.01, showing little disparity in false positive rates.

Overall, the baseline accuracy was statistically more significant than the proportional sampling but not statistically significant compared to the other sampling methods. For macro-F1, proportional sampling was statistically significantly better than other sampling methods. For the EOD, PC-ADASYN was statistically better than other sampling methods, while the no-sampling method SP was statistically significantly better than other sampling methods. Additionally, DI was notably low across most methods except proportional sampling, reflecting the difficulty in ensuring balanced positive outcomes, and PE remained

low overall, with proportional sampling showing the highest disparity in false positive rates. Also, for both DI and PE, the low scores across all groups show the contradicting nature of the fairness metrics and how difficult they are to optimize.

4.2 German Credits Dataset Results

The results of the experiments on the German Credit dataset also show varying impacts of each sampling strategy, particularly regarding fairness and accuracy, as shown in Table 4.2. The no-sampling method achieved an accuracy of 0.72, a macro-precision of 0.66, a macro-recall of 0.61, and a macro-f1 of 0.62 but exhibited significant bias in its prediction, with an EOD of 0.88, indicating a substantial disparity in error rates between groups in the protected attributes. However, the SP of 0.07 shows a low disparity between positive outcomes between protected and unprotected groups, contradicting the EOD. The baseline DI was 0.22, highlighting a severe imbalance in the ratio of positive outcomes across groups, while the PE was 0.18, reflecting a moderate disparity in false positive rates between protected and unprotected groups.

The oversampling method maintained accuracy while improving the precision to 0.67, recall to 0.69, and macro F1 to 0.68 and notably reducing EOD to 0.37, showing improvement in fairness. However, this came at the cost of increased SP to 0.35, highlighting a potential trade-off between different fairness measures. DI improved to 0.44, suggesting a better balance in positive outcomes between groups but still reflecting disparity. PE increased to 0.32, indicating that false positive disparities across groups had worsened slightly with this method.

Proportional sampling reduced accuracy slightly to 0.68, a precision of 0.69, and a recall of 0.68, giving this method the best F1-score of 0.69. It also lowered EOD to 0.32, suggesting it effectively balances prediction quality with fairness. The SP of 0.29 reflects a relatively low disparity in positive outcomes between groups, while DI increased to 0.59, showing a better

Table 4.2: Results of the sampling methods on the German Credit dataset with 95% confidence intervals. **Bolded** results indicate statistical significance.

Metric	No Sampling	Over-sample	Prop. Sample	PC-SMOTE	PC-ADASYN
Accuracy	0.72 ± 0.03	0.72 ± 0.05	0.69 ± 0.04	**0.73** ± 0.05	0.72 ± 0.03
Precision	0.66 ± 0.03	0.69 ± 0.02	0.69 ± 0.02	0.70 ± 0.02	**0.76** ± 0.02
Recall	0.61 ± 0.02	0.68 ± 0.02	**0.69** ± 0.02	0.55 ± 0.02	0.46 ± 0.02
Macro F1	0.62 ± 0.06	0.68 ± 0.04	**0.69** ± 0.07	0.55 ± 0.09	0.48 ± 0.11
EOD	0.88 ± 0.10	0.37 ± 0.16	0.32 ± 0.26	0.15 ± 0.09	**0.13** ± 0.02
SP	0.07 ± 0.01	0.35 ± 0.11	0.29 ± 0.10	0.10 ± 0.02	**0.06** ± 0.00
DI	0.22 ± 0.19	0.44 ± 0.15	0.59 ± 0.21	**0.00** ± 0.09	0.02 ± 0.03
PE	0.18 ± 0.11	**0.32** ± 0.11	0.21 ± 0.19	0.13 ± 0.12	0.02 ± 0.02

balance in outcome distributions. However, PE increased to 0.21, showing some remaining disparity in false positives, although the improvement in EOD suggests a reduction in bias in true positive rates.

PC-SMOTE improved accuracy to 0.73 and precision to 0.70 but a lower recall to 0.55 and macro F1 of 0.55. The model's fairness improved significantly over the baseline, with an EOD of 0.15 and a moderate SP of 0.1. In terms of DI, this method performed the worst with a value of 0.00, indicating a severe imbalance in prediction in the ratio of positive outcomes across groups. PE shows an improvement to 0.13, showing a reduction in false positive disparities.

PC-ADASYN yielded a similar accuracy to the baseline at 0.72 and an improved precision of 0.76, albeit with the lowest recall of 0.46 and macro F1 of 0.48, suggesting a potential trade-off in precision and recall and possibly a very small F1 score for one of the

classes. However, the model exhibited the best fairness performance with an EOD of 0.13 and SP of 0.06. DI improved slightly to 0.02, indicating that this method still struggled with a positive prediction ratio across each group. PE was the lowest at 0.02, suggesting that this method achieved the least disparity in false positive rates across protected groups.

Overall, the results show that the accuracy of no sampling was not statistically significantly different compared to other sampling methods except for proportional sampling. For the macro-f1, the no-sampling method result is statistically significant compared to PC-SMOTE and PC-ADASYN but not oversample or proportional sampling. For EOD, the results of all the sampling methods were statistically significant compared to the no-sampling method. Additionally, the DI of the no-sampling method is not statistically significant compared to the oversampling and proportional sampling method. PE of the no-sampling method was not statistically significant to PC-ADASYN and PC-SMOTE, reflecting the strengths of these methods in addressing different aspects of fairness.

4.3 Compass Dataset Results

Table 4.3 shows the results of our experiments with the COMPAS dataset. These results reveal significant variations in model performance across the different sampling strategies. The no-sampling method achieved an accuracy, macro-precision of 0.89, macro-recall of 0.89, and macro F1-score of 0.89 but showed higher disparities in fairness metrics, with an EOD of 0.39 and a SP of 0.29. Additionally, the DI was 0.58, indicating a moderate imbalance in the ratio of positive outcomes between protected and unprotected groups, and the PE was 0.39, suggesting a noticeable disparity in false positive rates across groups.

The application of over-sampling slightly improved accuracy to 0.90, precision to 0.91, recall to 0.91, and macro F1 to 0.90 while also improving fairness notably, decreasing EOD to 0.26. This suggests effectiveness in reducing outcome disparities without compromising SP, which decreased slightly to 0.25. However, DI remained relatively unchanged at 0.51,

Table 4.3: Results of the sampling methods on the COMPAS dataset with 95% confidence intervals

Metric	No Sampling	Over-sample	Prop. Sample	PC-SMOTE	PC-ADASYN
Accuracy	0.89 ± 0.04	0.90 ± 0.03	0.90 ± 0.05	**0.91** ± 0.02	0.91 ± 0.02
Precision	0.90 ± 0.03	0.91 ± 0.03	0.91 ± 0.02	**0.92** ± 0.03	0.91 ± 0.03
Recall	0.90 ± 0.03	0.91 ± 0.03	0.90 ± 0.02	**0.92** ± 0.03	0.91 ± 0.03
Macro F1	0.89 ± 0.04	0.90 ± 0.05	0.91 ± 0.04	**0.91** ± 0.02	0.91 ± 0.03
EOD	0.39 ± 0.15	0.26 ± 0.14	**0.26** ± 0.12	0.30 ± 0.11	0.30 ± 0.13
SP	0.29 ± 0.17	**0.25** ± 0.07	0.36 ± 0.10	0.47 ± 0.19	0.47 ± 0.21
DI	0.58 ± 0.27	0.51 ± 0.30	0.51 ± 0.16	**0.39** ± 0.19	0.40 ± 0.21
PE	0.39 ± 0.12	0.31 ± 0.23	**0.27** ± 0.16	0.30 ± 0.11	0.30 ± 0.13

indicating that the imbalance in positive outcomes persisted. PE improved slightly to 0.31, reducing the disparity in false positive rates.

Conversely, while boosting accuracy, precision, recall, and macro F1 to 0.90, 0.91, 0.90, and 0.91, respectively, proportional sampling also achieved an EOD of 0.26, improving it over the baseline. However, it recorded a higher SP of 0.36, indicating a potential increase in the disparity of positive outcomes across groups. DI remained at 0.51, similar to over-sampling, suggesting that proportional sampling still exhibits some bias in the ratio of positive outcomes between each group. PE improved to 0.27, reflecting a slight reduction in false positive rate disparities.

PC-SMOTE and PC-ADASYN had identical scores in accuracy, macro F1, and SP. The precision and recall were slightly different from each other, but both yielded precision and recall that surpassed the baseline. PC-SMOTE yields a precision and recall of 0.91 each,

while PC-ADASYN yields a precision and recall of 0.92. These methods also managed to maintain fairness improvements with an EOD of 0.30. However, both methods also increased SP to 0.47, indicating a greater disparity in positive outcomes between groups. Regarding DI, PC-SMOTE achieved 0.39, and PC-ADASYN reached 0.40, both showing some biases in balancing positive outcomes compared to the baseline. PE improved to 0.30 for both methods, indicating a reduction in false positive rate disparities compared to the baseline.

Overall, the results show that our sampling methods' accuracy, macro-F1, and EOD were statistically significantly better than the no-sampling method. However, the DI of no-sampling was statistically significant compared to other sampling methods, while the PE of all sampling methods was statistically significant compared to the no-sampling sampling. The variation in the fairness metrics results across the sampling methods reinforces how challenging it is to fully optimize these metrics, highlighting the difficulty of balancing multiple aspects of fairness simultaneously.

4.4 Bank Credit Default Dataset Results

Table 4.4 shows the results of our experiments with the Bank credit default dataset. The no-sampling method achieved decent performance with an accuracy of 0.82, macro-precision of 0.75, macro-recall of 0.65, and a macro-F1-score of 0.67. The fairness metrics reveal relatively low disparities, with EOD at 0.09 and SP at 0.04. However, DI was 0.67, indicating that protected groups received positive outcomes 67% as often as unprotected groups. PE was also low at 0.03, suggesting minimal bias in misclassification.

The oversampling approach slightly improved model fairness, with a small improvement in fairness. The accuracy remains the same at 0.82, precision at 0.75, recall at 0.65, and the F1-score also remained at 0.67. The fairness metrics show stability, with EOD remaining at 0.09 and SP at 0.04. However, Disparate Impact increased to 0.69, suggesting a slight improvement in the ratio of positive outcomes between groups. PE increased to 0.09, which

Table 4.4: Results of the sampling methods on Bank Credit dataset with 95% confidence intervals

Metric	No Sampling	Oversample	Prop Sample	PC-SMOTE	PC-ADASYN
Accuracy	0.82 ± 0.05	**0.82** ± 0.02	0.68 ± 0.14	0.82 ± 0.10	0.82 ± 0.09
Precision	0.75 ± 0.04	0.75 ± 0.04	0.69 ± 0.03	**0.77** ± 0.02	0.77 ± 0.03
Recall	0.65 ± 0.03	0.65 ± 0.03	**0.68** ± 0.03	0.63 ± 0.03	0.63 ± 0.03
F1-score	0.67 ± 0.03	0.67 ± 0.04	**0.68** ± 0.02	0.65 ± 0.04	0.66 ± 0.02
EOD	0.09 ± 0.03	0.09 ± 0.01	**0.05** ± 0.02	0.08 ± 0.03	0.09 ± 0.01
SP	0.04 ± 0.02	0.04 ± 0.01	0.07 ± 0.06	**0.03** ± 0.01	0.03 ± 0.02
DI	0.67 ± 0.05	0.69 ± 0.09	**0.89** ± 0.10	0.75 ± 0.08	0.72 ± 0.10
PE	0.03 ± 0.02	0.09 ± 0.02	**0.04** ± 0.01	0.07 ± 0.04	0.08 ± 0.07

may indicate a minor rise in false positive disparities across groups.

Proportional sampling decreased the model's accuracy to 0.68 and precision to 0.69 but slightly improved recall to 0.68 and the F1 Score to 0.68. Interestingly, this method resulted in better fairness, as indicated by the lower EOD of 0.05, suggesting a reduction in the difference between true and false positive rates between groups. However, SP increased to 0.07, and DI significantly improved to 0.89, indicating that positive outcomes are much more evenly distributed between groups. PE remained low at 0.04, reflecting minimal disparity in false positive rates.

The PC-SMOTE method maintained an accuracy of 0.82, similar to the baseline, while improving to 0.77, but recall dropped to 0.63, and F1-score dropped slightly to 0.65. Regarding fairness, EOD dropped to 0.08, showing a decrease in balancing true positive rates between groups, while SP decreased to 0.03. DI increased to 0.75, indicating a better

ratio of positive outcomes between groups, and PE slightly increased to 0.07, reflecting some disparity in false positive rates.

The PC-ADASYN method also maintained an accuracy of 0.82, an improved precision of 0.77, a lower recall of 0.63, and an F1 score of 0.66, which was close to the baseline. The fairness metrics performed similarly to the no-sampling method with an EOD of 0.09 and SP of 0.03. DI improved slightly to 0.72, reflecting better balance in positive outcomes across groups, while PE increased to 0.08, indicating some rise in the false positive disparity.

Overall, the results show that the accuracy of the no-sampling method was not statistically significant compared to other sampling methods except proportional sampling, while the macro F1 is statistically significant compared to other sampling methods except proportional sampling. The EOD, on the other hand, showed statistical significance in proportional and PC-SMOTE. The DI of the no-sampling method was not statistically significant compared to other sampling methods, while PE results showed that the no-sampling method was statistically significant compared to other sampling methods.

4.5 CDC Diabetes Dataset Results

The results of the experiments on the CDC Diabetes dataset, presented in Table 4.5, reveal notable variations in model performance across the different sampling strategies. The no-sampling method achieved a strong accuracy of 0.86, serving as a high baseline for comparison, a macro-precision of 0.72, a macro-recall of 0.56, and a macro-F1 of 0.57. Also, the model exhibited good fairness, with an EOD of 0.09, indicating a moderate disparity in error rates between protected and unprotected groups. The SP was very low at 0.01, showing minimal differences in the proportion of positive outcomes across groups. However, this may signal that the model was not making enough positive predictions for protected groups. The DI was 0.67, indicating an imbalance in positive outcomes between groups, while PE of 0.01 suggests minimal disparity in false positive rates. These results show that the baseline model

Table 4.5: Results of CDC Diabetes Dataset with 95% confidence intervals

Metric	No Sampling	Oversample	Prop. Sample	PC-SMOTE	PC-ADASYN
Accuracy	0.86 ± 0.06	0.86 ± 0.09	0.74 ± 0.10	0.86 ± 0.06	**0.86** ± 0.03
Precision	0.72 ± 0.03	0.72 ± 0.03	**0.73** ± 0.03	0.72 ± 0.02	0.71 ± 0.02
Recall	0.56 ± 0.03	0.54 ± 0.03	**0.72** ± 0.03	0.54 ± 0.03	0.53 ± 0.03
F1-score	0.57 ± 0.07	0.55 ± 0.04	**0.71** ± 0.02	0.54 ± 0.08	0.52 ± 0.08
EOD	0.09 ± 0.10	**0.08** ± 0.08	0.20 ± 0.07	0.12 ± 0.08	0.12 ± 0.03
SP	0.01 ± 0.05	**0.00** ± 0.02	0.15 ± 0.05	0.01 ± 0.02	0.02 ± 0.04
DI	0.67 ± 0.04	**0.83** ± 0.09	0.78 ± 0.08	0.31 ± 0.10	0.00 ± 0.05
PE	0.01 ± 0.02	**0.00** ± 0.01	0.20 ± 0.10	0.00 ± 0.02	0.01 ± 0.03

performed well in terms of both accuracy and fairness.

The over-sampling method maintained the baseline accuracy of 0.86 and a precision of 0.72 but slightly dropped the recall to 0.54 and the macro F1-score to 0.55, suggesting possible overfitting and difficulty in balancing precision and recall. Regarding fairness, the EOD slightly improved to 0.08, indicating a marginal reduction in disparity between groups. Also, SP remained at 0.00, suggesting no change in the distribution of positive outcomes between groups. Notably, DI increased to 0.83, reflecting a better balance of outcomes, but PE stayed very low at 0.00, showing no significant difference in false positive disparities.

The proportional sampling method led to a substantial drop in accuracy to 0.74 but significantly improved the precision to 0.73, recall to 0.72, and the macro F1-score to 0.71, suggesting that the model could balance precision and recall more effectively. However, the EOD increased to 0.20, indicating that the model struggled to balance error rates between groups. Similarly, SP increased to 0.15, signaling a greater disparity in the distribution of

positive outcomes. DI was slightly increased at 0.78, showing improvement in the ratio of positive outcomes across groups. PE increased to 0.20, reflecting a higher disparity in false positive rates, suggesting that proportional sampling, while improving precision and recall, had mixed effects on fairness.

PC-SMOTE maintained the baseline accuracy of 0.86 and baseline precision of 0.72 but saw a drop in the recall to 0.54 and the macro F1-score to 0.54, indicating challenges in balancing precision and recall. The EOD increased slightly to 0.12, suggesting the method was less effective in reducing disparities in error rates compared to other methods. SP remained very low at 0.01, and DI dropped significantly to 0.31, indicating a substantial imbalance in the ratio of positive outcomes between groups. PE remained at 0.00, showing no improvement in false positive disparities.

The PC-ADASYN method showed a similar accuracy to the baseline at 0.86 but dropped precision to 0.71, recall to 0.53, and had the lowest F1-score of 0.52, suggesting challenges in precision and recall. In terms of fairness, EOD was similar to PC-SMOTE at 0.12, showing moderate disparities in error rates across groups. SP was slightly better at 0.02, but DI dropped to 0.00, indicating a severe imbalance in positive outcomes between groups. PE remained low at 0.01, reflecting a minimal disparity in false positive rates.

Overall, the results show that the accuracy of the no-sampling method was not statistically significant compared to other sampling methods except proportional sampling. The macro-F1 of the no-sampling method was statistically significant compared to other sampling methods except proportional sampling. The EOD of the no-sampling method showed statistical significance in all the sampling methods except proportional sampling. The DI of the no-sampling method was statistically significant compared to other sampling methods except oversample and proportional sampling, while PE results showed that the no-sampling method was statistically significant compared to other sampling methods except proportional sampling.

CHAPTER FIVE

ANALYSIS AND DISCUSSION

Results of the experiments on the five datasets strongly support that protected-category sampling can enhance model fairness, often without significantly compromising prediction accuracy. In some cases, improvement in accuracy and macro-F1 were also demonstrated.

5.1 Comparing Fairness vs performance

In this section, we explained the performance of each sampling method and each dataset on both classification and fairness metrics performance.

5.1.1 Adult Income Dataset

Focusing on the Adult Income dataset results, PC-SMOTE and PC-ADASYN demonstrated notable EOD improvements and maintained moderate SP levels. The efficacy of these methods can largely be attributed to their sophisticated interpolation techniques. For example, visually examining the decision trees generated with no-sampling and PC-ADASYN provides insightful contrasts. Examples from a single representative fold are shown in Figures 5.1 and 5.2, respectively. The decision tree learned without sampling selected its root with a feature closely associated with protected attributes, thus acting as a proxy attribute. This led to pronounced prediction bias, as reflected in the EOD. Conversely, the decision tree trained on data generated using PC-ADASYN began with a feature that generalized predictions very well and mitigated bias, as evidenced by a notable enhancement in model fairness and a higher Gini impurity, indicating a purer initial split.

Comparing over-sampling and proportional sampling, the methods' approaches to augmenting sample size by duplicating existing data rows were straightforward and did not yield substantial improvements in EOD. This outcome makes sense since these methods

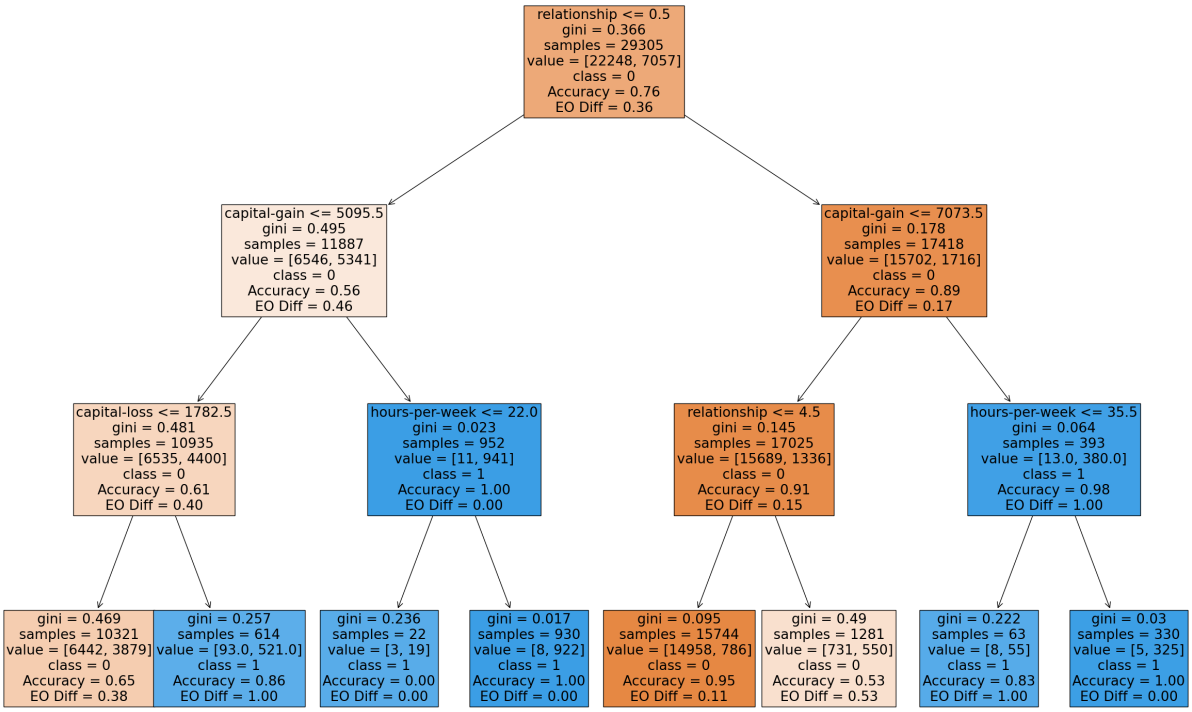

Figure 5.1: Example decision tree trained on Adult Income with no sampling

tend to replicate existing biases, which can potentially exacerbate fairness issues rather than alleviate them. This is evident, in particular, when considering the Adult Income dataset, where the classes are highly imbalanced. These naïve replication strategies lack the interpolation capacity of PC-SMOTE and PC-ADASYN to adjust samples near decision boundaries, which is crucial for mitigating the bias in the dataset. In contrast, the interpolation strategies used by PC-SMOTE and PC-ADASYN expand the dataset and enhance its diversity. This is particularly effective for samples near decision boundaries, where slight shifts in the features can affect the fairness of predictions significantly. By interpolating between samples, PC-SMOTE and PC-ADASYN effectively move these boundary samples towards more equitable regions of the feature space, thus directly

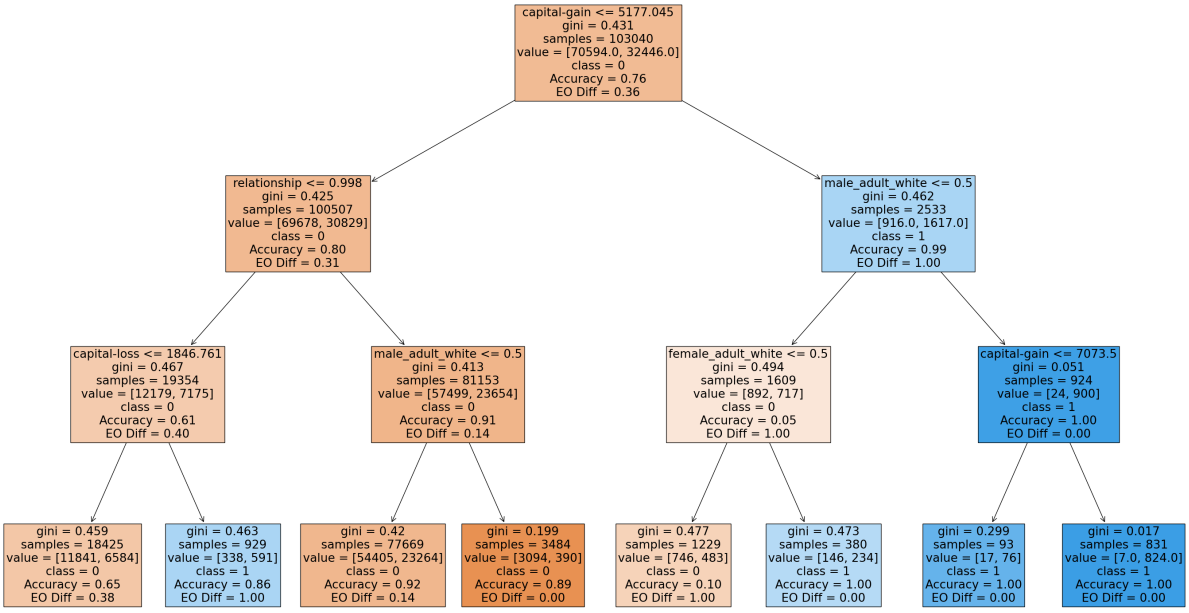

Figure 5.2: Decision tree of PC-ADASYN on Adult Income

confronting and reducing bias more effectively than methods that increase sample volume without altering data structure. The class generation function (Algorithm 3.4) also helps improve the overall class distribution.

From the DI and PE perspectives, we can see a contradicting result that shows the difficulty of optimizing for multiple fairness metrics. Specifically, the DI shows the results are biased while the PE result shows that the result is unbiased, this comes as a result of the conflicting methods each fairness metric is using to measure unfairness. The DI measures the rate of positive outcomes received by the unprivileged group compared to the privileged group. In contrast, PE measures the difference between the false positive rate between each group. The results of DI exhibiting bias while PE exhibiting fairness show the difficulty of optimizing for multiple metrics, especially conflicting metrics like DI and PE.

5.1.2 German Credit Dataset

In examining the German Credit dataset results, we observed a trend similar to what was noted in the Adult Income dataset: the no-sampling method had a very high bias regarding EOD. The low SP of 0.07 indicates minimal disparity in the positive prediction rates between the groups. However, this is not a good way of measuring fairness since the favored group has more samples than the unfavored ones. This skews the SP calculation since it only counts positive decisions in each group, which are influenced by sample size. Regarding the DI, the results show that all the sampling methods exhibit some biases. This is expected as the DI, which measures the ratio of positive outcomes for each group, and since the dataset is imbalanced, the number of positive outcomes in one group is skewed compared to the other group. For PE, which measures the false positive rate, we can see that our sampling methods (PC-SMOTE and PC-ADASYN) outperform the baseline method. One takeaway is the importance of employing multiple fairness metrics to view a model's impact on all stakeholders comprehensively. For over-sampling, we saw an improvement in EOD with a similar SP; this shows that increasing the number of samples for each of the multicategory's protected attributes improves the fairness with respect to EOD. In addition, the updated SP reflects what it will look like to have an equal number of samples for each multi-category, unlike in the baseline, where the favored group has five times more samples than the unfavored group. The DI also improves over the baseline because the oversampling increases the number of positive outcomes across the protected groups, while the PE decreases over the baseline because as the number of positive outcomes increases, the false positive rate also tends to increase because we now have many positive samples. The results of the metrics still further emphasize the existing contradicting relationship between these two metrics and also among fairness metrics.

The accuracy of proportional sampling drops because the baseline number of samples selected after hyperparameter tuning was insufficient for the model to generalize the

unsampled test set, leading to overfitting. The overfitting was confirmed by considering the training accuracy. Interestingly, the model does not trade accuracy for recall like other models, giving proportional sampling the highest macro-F1. Regarding fairness, we found an increase in EOD compared to baseline and over-sampling. This arises because the same number of samples now represents each multi-category as the favored category. This can improve the model fairness because the model now has a bigger picture of categories and makes better predictions and, ultimately, fairer decisions. The DI also increased because we made a target sample number for each category, which tends to decrease the number of negative samples in the protected category while potentially increasing the positive samples in the unprotected category. Also, we see an increase in the PE, which shows that increase in the number of false positive rates, this occurs because as we increase the number of positive samples in the dataset, there is a tendency of many false positive samples.

PC-SMOTE and PC-ADASYN play pivotal roles in significantly reducing bias in model predictions. This consistency confirmed the robustness of these methods across different datasets. Notably, neither method compromises accuracy, while both enhance fairness, illustrating their effectiveness in handling the trade-offs typically associated with predictive modeling. These results demonstrate the strong interpolation capabilities inherent in PC-SMOTE and PC-ADASYN. These methods effectively reallocate samples within the feature space, especially moving those in underprivileged regions from negative to more positive decision boundaries. Such adjustments are crucial in mitigating biased outcomes and promoting equity in automated decision-making processes.

The macro F1 in both models drops compared to the baseline because a higher number of samples is required for the model to perform better on generalization, which this dataset does not support. Specifically, the dataset has 700 samples for class 0 and 300 samples for class 1, which means the test set only has 30 samples for class 1. This small number of samples forced both models to trade recall for precision in class 1. Notably, we saw a low

recall for class 1, ultimately leading to a low F1 score for class 1. Since macro F1 averages the two F1 scores and treats them equally, this affects the performance of both models in macro F1. Overall, the two models yield a fairer model with good accuracy compared to the baseline and the other two sampling methods, except in the case of DI, where the two models drop. This happens because the metric only focuses on positive outcomes, and this dataset has very few positive outcomes for each group. For the DI, the number of positive outcomes stays the same because the sampling method did not specifically replicate the positive samples in the existing dataset, but it interpolates, which can potentially increase both positive and negative classes. The DE of both models improves over the baseline because the methods handle fairness better, especially in the area of false positives and false negatives, as we saw in EOD.

5.1.3 COMPAS Dataset

The COMPAS dataset's evaluation further validates our sampling methods' effectiveness. The distinct patterns that emerged align with those observed in the Adult Income and German Credit datasets, underscoring the robustness of our findings. Notably, over-sampling and proportional sampling techniques demonstrated substantial improvements in EOD and accuracy, while over-sampling also notably improved in SP. This improvement is likely due to the unique composition and balance within the COMPAS dataset, unlike the other datasets in which the classes are imbalanced. The success of over-sampling and proportional sampling in this context can be attributed to the balanced nature of the dataset, which allows repeated duplication of existing rows (sampling techniques employed by these methods) to enhance the dataset without introducing a significant skew towards any particular class. This method effectively augments the representation of all classes and the protected attributes in a balanced form, making these techniques particularly effective for datasets where the feature domains contribute equally to predictions and where initial

class distributions do not suffer from severe imbalance. This can further be verified from their macro F1 as none of the models trade precision for recall. The improvement of SP in over-sampling can be attributed to the higher number of samples in over-sampling in comparison to proportional sampling.

Regarding DI, over-sampling and proportional sampling methods showed a moderate decline over the baseline. Specifically, the DI for over-sampling and proportional sampling decreased to around 0.51, indicating that replication strategies that both sampling methods rely on did not improve the number of positive outcomes. This reduction in DI also suggests that these sampling methods slightly introduce some skewness in the number of positive outcomes across different groups. However, it is important to note that both methods (no-sampling and oversampling) still fall short of achieving perfect positive prediction, likely due to the challenge of optimizing multiple fairness metrics simultaneously. Furthermore, the nature of the COMPAS dataset may inherently contain lower positive outcome rates for certain protected groups, making it difficult to fully balance DI. When analyzing PE, which measures false positive rate disparities, over-sampling, and proportional sampling methods improved upon the no-sampling method, but these gains were less substantial. This shows that the model's positive results are truly classified as positive, unlike in the baseline, where we have more false positive results. These metrics indicate that while the models performed well in terms of overall accuracy, PE, and EOD, they struggled in the areas of DI and SP, both of which count positive predictions. The difficulty in simultaneously optimizing for multiple fairness metrics, especially in skewed datasets, is evident here.

For PC-SMOTE and PC-ADASYN, these algorithms showed an improvement over the baseline in both accuracy and EOD, as well as PE. However, the DI for both methods remained lower than the no-sampling method, suggesting a greater disparity in positive outcomes between groups compared to the over-sampling and proportional sampling methods. This could be due to the inherent difficulties these algorithms face in balancing

between optimizing for multiple fairness metrics. Also, this could be because the distribution of positive outcomes in the protected and unprotected groups are significantly different, and the algorithm did not trade more positive predictions for false positives, as we can deduce from the PE of the no-sampling method.

A notable trend in the results was the large drop in SP for these methods, which can be attributed to the new label that skewed the dataset towards the negative class. The drop in SP is also reflected in the DI, as the skewed dataset made it difficult for these algorithms to balance positive outcomes between groups. Furthermore, the lower positive class rate in one of the groups may have exacerbated these issues, leading to the observed disparities in fairness metrics. These results also highlight the complexity of optimizing for multiple fairness metrics simultaneously. Each fairness metric targets a different aspect of fairness in the model, and improving one can often lead to a trade-off with others. The inherent characteristics of the dataset, such as skewed class distributions or lower positive outcome rates for certain groups, can further complicate this optimization process and the choice of fairness metrics.

5.1.4 Bank Default Credit Dataset

The results of the Bank Default Credit dataset, as shown in Table 4.4, indicate varying performances of the different sampling methods. The no-sampling baseline achieved a reasonable accuracy, a moderate F1-score, and moderate performances across fairness metrics. The EOD of 0.09 indicates a low difference in TPR and FPR between groups, while the SP of 0.04 suggests the positive prediction of the protected and unprotected groups is also similar. Moreover, DI reflects a disparity in positive outcomes between protected and unprotected groups, where protected groups receive 67% as many positive outcomes. Similarly, the PE of 0.03 highlights a minimal gap in false positive rates between groups. These metrics suggest that the model performs well in accuracy and fairness except in DI.

This result is unsurprising as the dataset is neither noisy nor missing, making it suitable for fairness evaluation.

The Oversampling method maintained the same accuracy and macro-F1 as the baseline method and slightly improved some fairness metrics, especially in DI. The EOD remained the same as no-sampling, indicating that the method did not substantially improve the balance of TPR and FPR across groups. However, DI increases, showing a marginal improvement in the ratio of positive outcomes between groups. The increment in DI can be attributed to the replication (or oversampling) of the dataset, which in turn produces more positive samples. Also, the increment can be attributed to the rising number of false positives, which can be confirmed in the PE scores, as we can deduce from the results that oversampling increases the number of false positives. Overall, we can see how the impact of many fairness metrics can affect the fairness results. Also, we can see the contradicting nature of different fairness metrics.

Proportional Sampling, however, showed different results, sacrificing accuracy in exchange for significant improvements in fairness metrics and F1-score. The EOD increased, meaning the unprotected groups achieved a fairer result compared to the baseline. This happens because the number of samples in the unprotected categories increases because of the sampling method, which in turn gives an unbiased representation of each group to the model. DI increased to 0.89, the highest among all methods, indicating a much more balanced distribution of positive outcomes across groups. Also, we can see the impact of proportional sampling in improving the number of positive samples in the unprotected group. The improvement in DI that we see in proportional sampling is better than that we saw in oversampling, as this method did not improve positive prediction at the expense of increased false positives as we saw in the oversampling. Additionally, SP increased slightly to 0.07, showing a moderate decrease in the parity of positive outcomes. This occurs because this sampling method eliminates some samples in the favored group, thereby losing some positive

predictions. The improvement in fairness, especially in DI and EOD, can be attributed to the method's ability to better balance the representation of groups, particularly in datasets where class distributions are not severely imbalanced. However, the drop in accuracy shows that proportional sampling may struggle with performance when working with complex feature interactions and when working with fewer samples than the test set.

PC-SMOTE and PC-ADASYN both showed similar results, maintaining the baseline accuracy while slightly decreasing macro-F1. This happens because both methods trade precision for recall. This is a result of the dataset's imbalance and the number of samples in their test set. The fairness metrics also see a slight increase, for example, in EOD. This occurs because the sampling strategy improves the overall fairness of the dataset. This further demonstrates the power of interpolation in these sampling strategies. The interpolation helps generate new samples that are fairer, which in turn improves the EOD of PC-SMOTE while keeping the same baseline for PC-ADASYN. We can also demonstrate that both models yield a better SP; this is unsurprising as it further reinforces the superior power of these sampling strategies and how they can improve the overall positive prediction, especially for the unfavoured group. The improvement in these positive predictions can be seen further in DI. Both sampling methods successfully improved the number of positive predictions in the dataset. This stems from the fact that both methods move the fairness boundary of the newly generated synthetic sample. The slight drops in the PE can be associated with the increase in the number of false positives in the dataset. This is also unsurprising as the number of false positives can increase as we increase the number of positive predictions. This can be attributed to three likely causes. First, the sampling methods made some false positive predictions. This can occur because our sampling methods generate and predict some negative samples as positive. The second reason can be attributed to the distribution of positive samples in the dataset. This means if the positive distribution in the dataset is skewed, the only way to improve positive distributions is at the expense of increasing false

positives. This is one of the reasons why SP and DI are not the best fairness metrics because they can be misleading. The third reason is the contradicting nature of the fairness metrics and the relationship between the dataset and fairness metrics. Before choosing any fairness metrics, there might be a need to understand the dataset characteristics, the impact any bias mitigation strategy will have on the dataset, and how this strategy will negatively or positively impact other fairness metrics.

Overall, in this dataset, Proportional Sampling stood out as the most effective method for improving fairness in terms of EOD and DI, despite the drop in accuracy, while PC-SMOTE and PC-ADASYN offered modest improvements in fairness metrics without significantly decreasing accuracy. Oversampling performed similarly to the baseline, maintaining stable accuracy but making only minor improvements in fairness metrics.

5.1.5 CDC Diabetes Dataset

The results of the experiments on the CDC Diabetes dataset show varied performances across the different sampling strategies. The no-sampling baseline achieved a reasonable accuracy of 0.86, with a moderate EOD of 0.09, which shows a low disparity between the false positive rate and the false negative rate of the favored and unfavored group. The F1-score is slightly low at 0.57, which suggests that there might be a trade-off between precision and recall in one of the two classes. A further examination later shows the recall for class 1 was specifically low, affecting the macro-F1 of class 1 and the overall macro-F1. The SP of 0.01 shows that the difference between positive prediction is very low and fair. Regarding the DI, the result shows the model exhibits some level of unfairness in the area of ratio of positive prediction between favored and unfavored groups. The PE of 0.01 shows a very low level of unfairness in the aspect of false positive rate between the protected and unprotected groups. Overall, the baseline set a good accuracy and a fair model, which can be attributed to the fair nature of the dataset.

The oversampling method maintained the same accuracy as the baseline method, but the F1 score slightly dropped to 0.55, indicating a far worse trade-off in precision and recall. Specifically, the recall of class 1 in the dataset is very low due to the extremely imbalanced nature of the dataset, thereby causing the algorithm to suffer misclassification. Regarding fairness, we can see an increase in all the fairness metrics over the baseline, which signifies that increasing the number of samples in each protected category helps in combating bias in the model prediction. The increase in DI shows that our sampling strategy generates more positive samples, especially for the unfavored group, thereby enhancing the model's ability to generate a fairer prediction for each group in comparison to the baseline. This sampling method also shows superiority in terms of DE as the algorithm did not increase positive prediction in DI at the expense of false positive prediction, as we see in PE. Rather, the PE drops, which shows that the method combats false positive prediction rather than exacerbating it.

Proportional sampling, in contrast, showed a trade-off with a significant reduction in accuracy to 0.74, though the F1-score improved to 0.71. Unlike all other sampling methods, this method did not trade precision for recall because it has an affinity to better balance its precision and recall due to its sampling strategy of selecting the target number of samples for each category. However, fairness metrics like EOD increased to 0.20, PE increased to 0.20, and SP increased to 0.15, indicating that proportional sampling struggled to reduce bias in its prediction. This can be attributed to the fact that the dataset is fair from the onset, and our sampling strategy introduces bias into the dataset rather than reducing bias. This bias likely came from the elimination of the majority group, which led to a loss of information. The increase in DI and the decrease in PE also show that this method likely trades more positive predictions for more false positive predictions. This further shows that this method struggles to reduce unfairness in the dataset, and it also shows how this sampling method can introduce bias if not carefully used.

Both PC-SMOTE and PC-ADASYN maintained the baseline accuracy of 0.86 but showed lower F1 scores at 0.54 and 0.52, respectively, suggesting challenges with precision and recall. This was visually confirmed from the result as the recall for diabetes (class 1) was really low due to the low sample number which was not enough for the sampling method to achieve good generalization. Regarding fairness, EOD and DI were low in comparison to the baseline because this dataset suffers from an extremely imbalanced nature, and the interpolation strategies employed by these sampling strategies struggle to improve the decision boundary in the dataset. The SP and PE were comparable to the baseline, which shows that these sampling methods did improve the positive prediction, especially for the unprotected group, which was a result of increasing the number of samples of the protected group.

Overall, oversampling is the best method because it maintains the same baseline accuracy while improving the EOD, SP, DI, and PE, although it slightly lowers the F1 score. PC-SMOTE and PC-ADASYN show competitive performances in accuracy, SP, and PE while slightly lowering EOD and DI significantly. This can be attributed to the imbalanced nature of the dataset and the hardness of optimizing for multiple fairness metrics.

5.2 Impact of Tree Depth on Fairness and Accuracy

In this thesis, the impact of decision tree depth on model performance was also investigated, specifically examining how variations in tree depth influence accuracy and EOD. Understanding the depth's effect provides insight into the effects ranging from underfitting to overfitting and helps identify the optimal complexity level at which both accuracy and fairness are maximized. Initially, the decision tree was allowed to grow without constraints to its full depth, which had an average depth of 30. The tree was then examined visually to deduce the maximum depth, excluding the non-splitting branches. To analyze the effects of tree depth systematically, the maximum depth of the trees was allowed to vary from 1 to

30. Each depth limit was evaluated using ten-fold cross-validation to ensure the robustness and generalizability of the findings.

For each configuration of tree depth, the accuracy and EOD were measured on the test set. Additionally, 95% confidence intervals were calculated for the metrics across the ten folds. This statistical analysis highlighted the depth at which the decision tree optimized both accuracy and fairness while also considering the underlying statistical bias-variance tradeoff. By doing so, it was possible to pinpoint the "sweet spot"—a delicate point where the decision tree maintains high predictive accuracy without compromising fairness, effectively countering the often-cited trade-off presented in previous literature. Figures 5.3–5.7 show the plots of accuracy and EOD against maximum depth for each of the five sampling methods on the Adult Income dataset.

Based on results such as those shown in Figures 5.3 and 5.4, there is a notable initial increase in accuracy as maximum depth increases for both the no-sampling and the over-sampling methods. However, both methods exhibit a decline in accuracy from a depth of 10 onwards, suggesting the onset of overfitting. Correspondingly, the EOD decreases sharply with increasing depth up to about depth 10, beyond which it stabilizes. This pattern indicates that while deeper trees initially improve fairness, they eventually reach a threshold beyond which no further gains are observed. Recalling that the fairness goal was to minimize EOD, a key observation is that setting the maximum depth between three and five corresponds to the optimal point for achieving high accuracy and maintaining low EOD.

When considering the results shown in Figure 5.5, the proportional sampling method continually increases accuracy with tree depth, peaking at a depth of about 26. Conversely, the EOD initially increases before decreasing and stabilizing at a depth of around 15. The wide confidence intervals observed in the EOD metric suggest significant variability in fairness outcomes. This finding underscores the importance of selecting a depth that minimizes variability in fairness while maximizing accuracy.

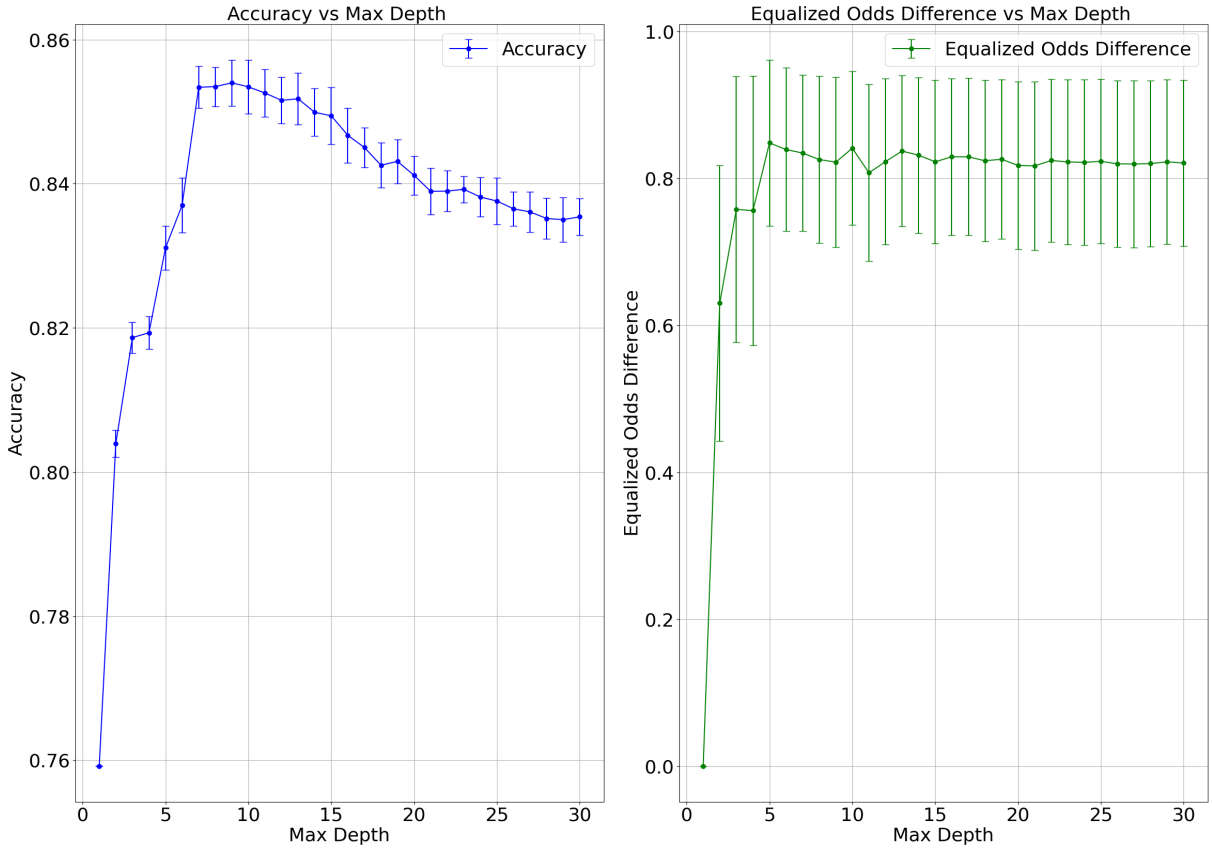

Figure 5.3: Plots of Adult Income using no sampling, showing accuracy and EOD with 95% confidence intervals against the maximum tree depth ranging from 1 to 30.

Figures 5.6 and 5.7, representing the results using PC-SMOTE and PC-ADASYN, respectively, exhibit slight downward trends in accuracy, which improve briefly before descending again—a pattern indicative of overfitting at greater depths. EOD metrics for these methods show initial stability at lower depths, surge at mid-level depths, and decline, suggesting complex interactions between synthetic sample generation and decision boundary delineation. Given these observations, a maximum depth of 3 was chosen for our experiments as it represented a "sweet spot" where both accuracy and EOD are optimized.

Given these results, one conclusion is to challenge the often presumed trade-off

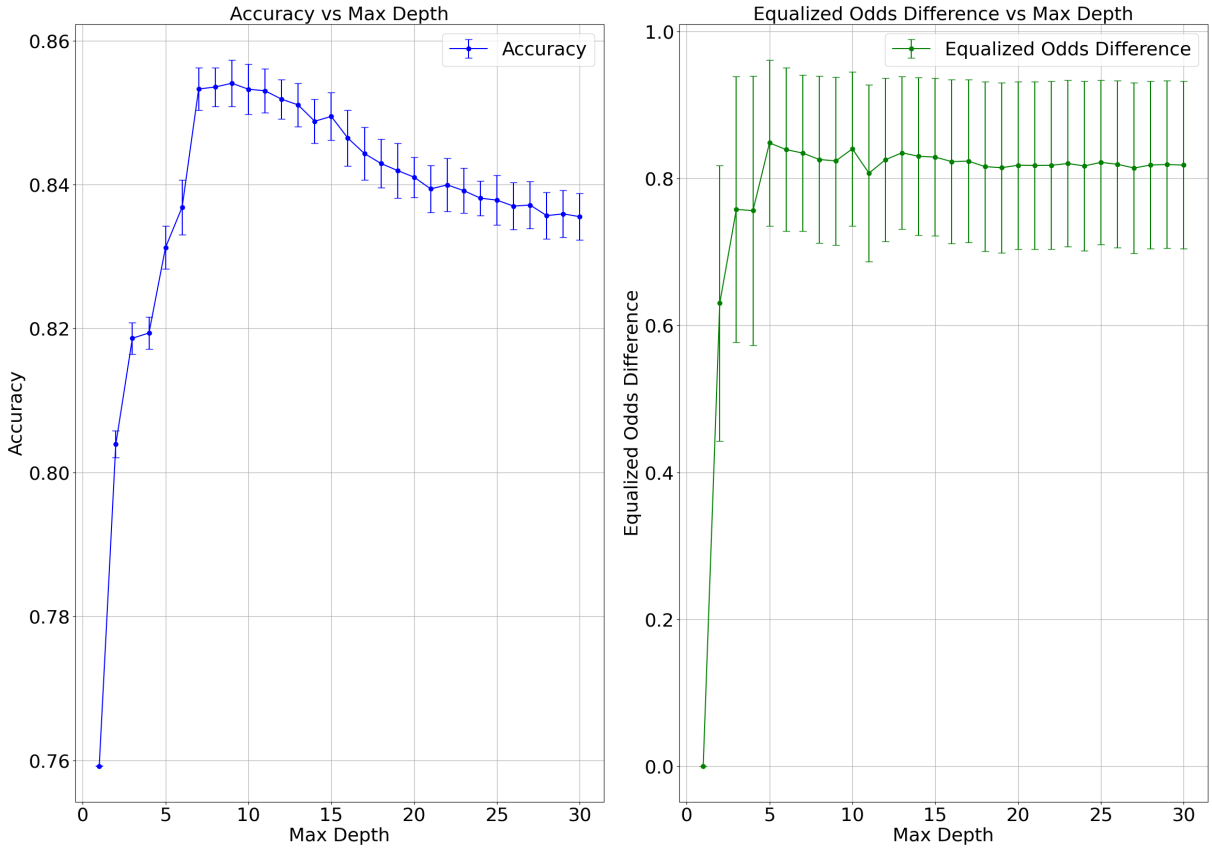

Figure 5.4: Plots of Adult Income using oversampling, showing accuracy and EOD with 95% confidence intervals against the maximum tree depth ranging from 1 to 30.

between accuracy and fairness by demonstrating that our PC-ADASYN method consistently outperforms baselines across all three datasets in terms of both accuracy and fairness. This finding is significant, as it suggests enhancing model fairness without sacrificing accuracy with appropriate sampling methods and model tuning is possible. However, our analysis also reveals scenarios where adjustments to model complexity, specifically the maximum depth of decision trees, can enhance accuracy at the expense of fairness, as indicated by increases in EOD. It is expected, however, that coupling sampling methods with in-processing methods such as fairness-based regularization may offset these effects. These

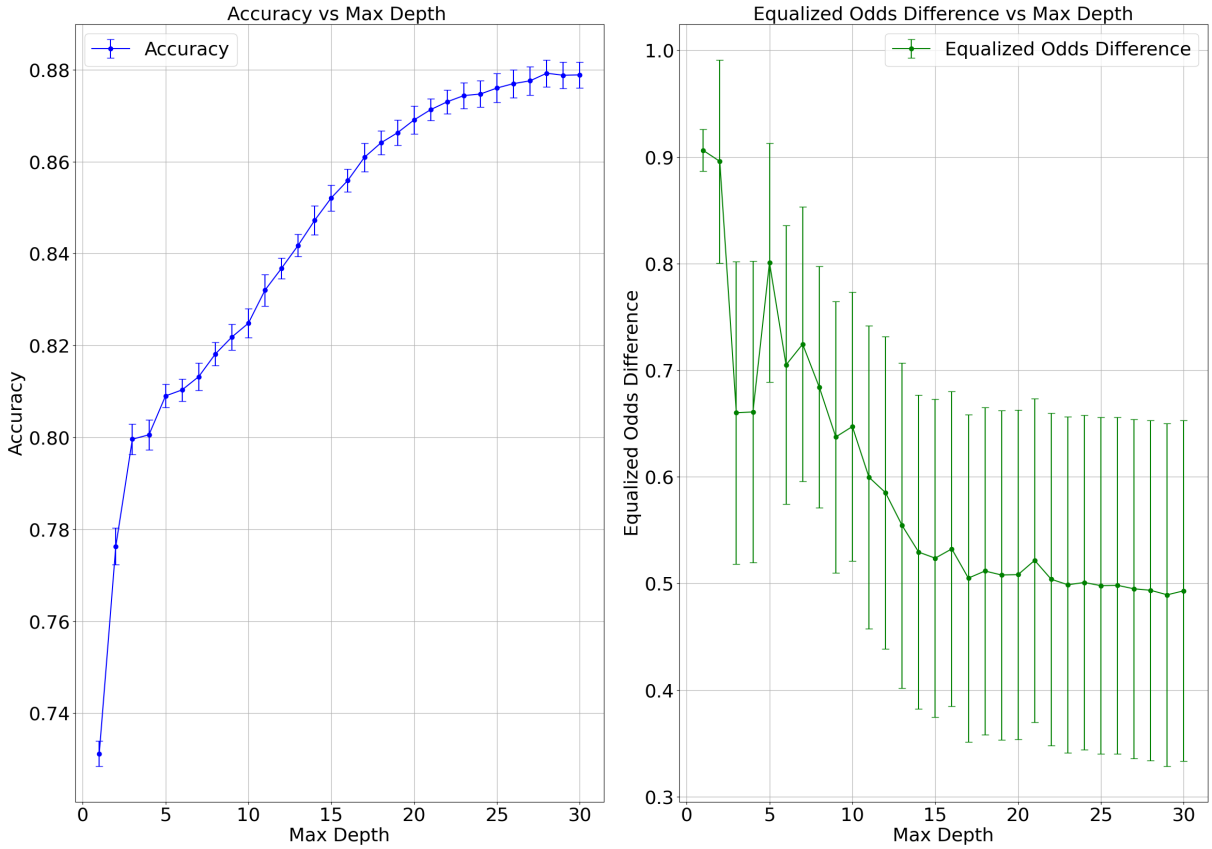

Figure 5.5: Plots of Adult Income using proportional sampling, showing accuracy and EOD with 95% confidence intervals against the maximum tree depth ranging from 1 to 30.

decisions highlight researchers' discretionary power in balancing model performance metrics depending on their study's specific objectives and constraints. The quantity of sample and time complexity is like every other sampling method. As the sample quantity increases, the time complexity increases, but overall, the sampling methods have the same time complexity as their underlying algorithms.

Moreover, our results underscore the complexities of simultaneously optimizing multiple fairness metrics. For instance, efforts to improve Statistical Parity (SP) by favoring more positive predictions for each protected group in the COMPAS dataset led to an inadvertent

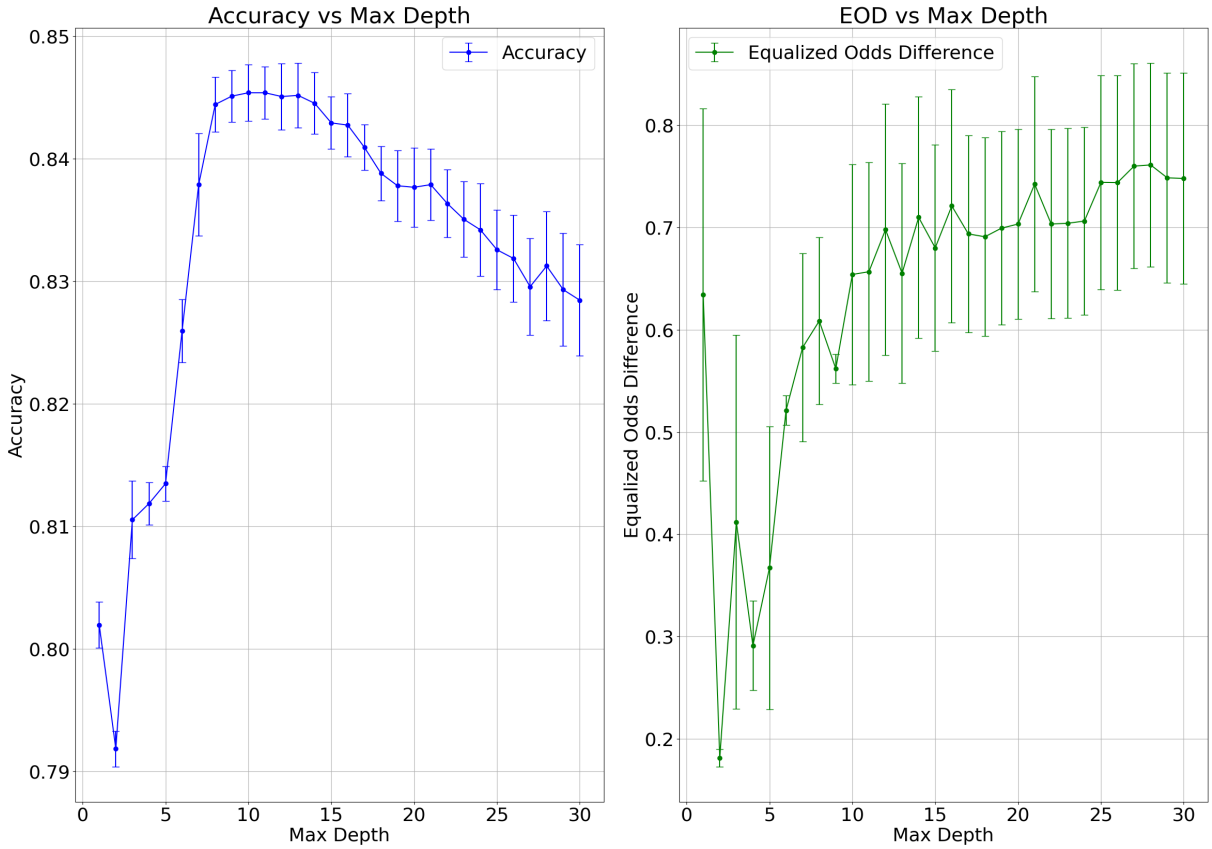

Figure 5.6: Plots of Adult Income using PC-SMOTE, showing accuracy and EOD with 95% confidence intervals against the maximum tree depth ranging from 1 to 30.

reduction in negative predictions. This shift adversely impacted the False Positive Rate (FPR), a component of EOD, thereby worsening the EOD metric as SP improved. This phenomenon illustrates the inherent mathematical tensions between fairness metrics, where optimizing one can affect another detrimentally. The COMPAS dataset, with its nearly balanced class distribution, provides a concrete example of how dataset characteristics can influence the behavior of fairness metrics. Optimizing SP in this context implies a skewed measurement of fairness, particularly where inherent differences exist between groups in protected attributes. This is supported by literature indicating that SP may not adequately

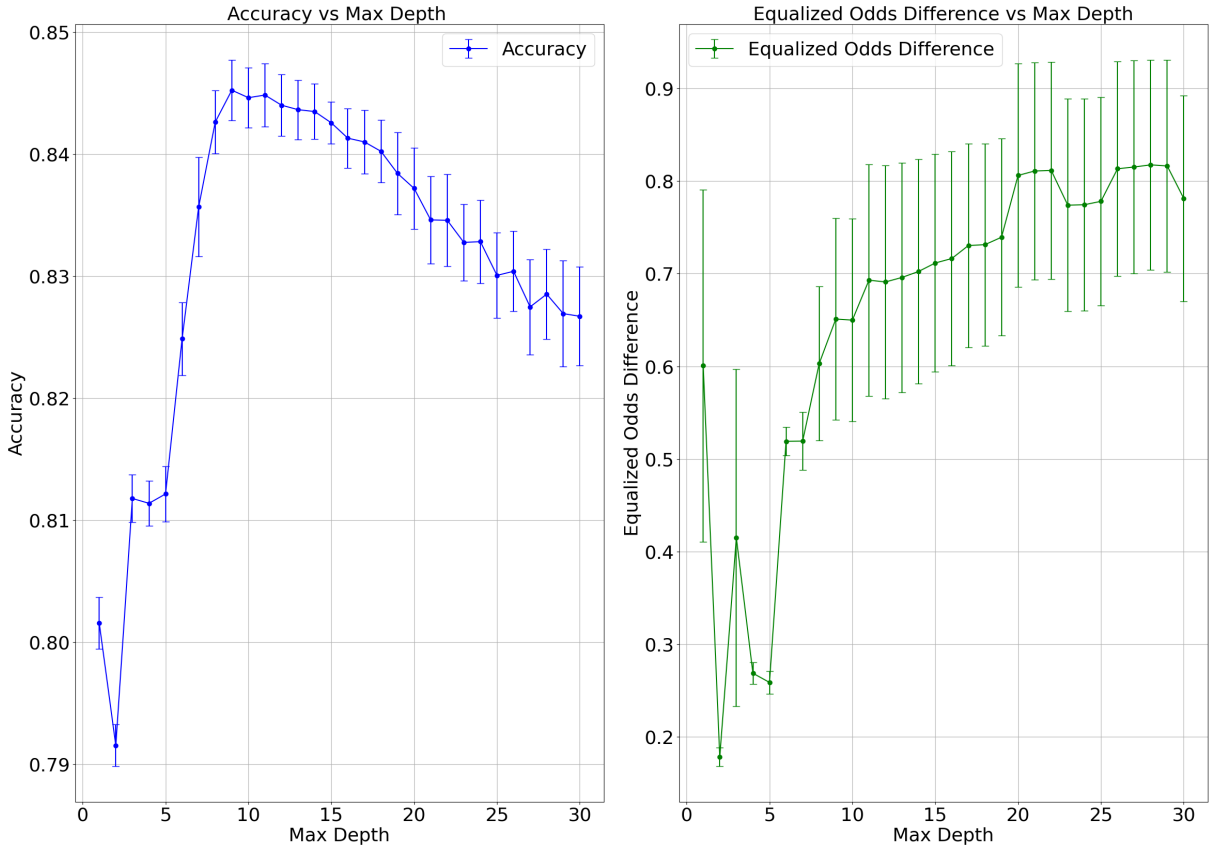

Figure 5.7: Plots of Adult Income using PC-ADASYN, showing accuracy and EOD with 95% confidence intervals against the maximum tree depth ranging from 1 to 30.

account for group differences, potentially leading to misleading conclusions about a model's fairness [19].

5.3 Impact of Our New Sampling Method on Class Label

In Chapter 3, we made a claim about how previous sampling methods are not suitable for our goal of sampling the protected category because they sample based on the class label, which is not our intent regarding protected category sampling. We devised new sampling methods that sampled these protected categories. Still, as we sampled these new multi-

categories, we also increased the number of samples in the dataset, subsequently increasing the number of examples associated with each class label. Still, we do not know how much impact each sampling method has on the class label itself. Hence, in this thesis, we conducted further experiments and statistical tests to determine how much the distribution of each class in our protected categories changed with respect to each sampling method. Understanding this, we inform ourselves of the impact of class labels on both prediction and fairness.

For this experiment, we used a single representation of our training set since our test set is unsampled. This experiment is designed to calculate the conditional class probabilities for each combined protected category (e.g., *adult_male_white*). The experiment calculated the conditional probabilities of $P(Y = 0|S = w)$ and $P(Y = 1|S = w)$ for a given set of true labels in the training set and corresponding protected attributes in the training set. This results in one probability for each class in our dataset for each protected attribute.

Upon extracting the above probabilities for all the sampling methods, we proceeded to compare the class distribution of each protected attribute in baseline (no-sampling) with the other four sampling methods. For each protected group, we used the KL divergence to measure the shift in the distribution from the class distribution of the baseline to the other sampling methods. The KL divergence measures how one probability of a distribution diverges from another probability distribution (i.e., to measure the divergence of distribution between the probability of each sampling to the no-sampling method). The KL divergence is calculated as follows:

$$D_{KL}(P||Q) = \sum P(i) \log \frac{P(i)}{Q(i)}$$

where P is the baseline distribution, and Q is the distribution after sampling. Upon calculating the KL divergence for each protected category, we selected a threshold of 0.05 as a statistical significance. The results in Table 5.1 through Table 5.5 show the probabilities of each class and each protected attribute of our five sampling methods on the adult income

Table 5.1: No-Sampling Conditional Probabilities of each class and each protected attribute of Adult Income Dataset

Protected_category	Class_0	Class_1
Female_Adult_Other	0.915	0.085
Female_Adult_White	0.842	0.158
Female_Young_Other	0.988	0.012
Female_Young_White	0.989	0.010
Male_Adult_Other	0.741	0.259
Male_Adult_White	0.625	0.375
Male_Young_Other	0.988	0.012
Male_Young_White	0.977	0.023

dataset.

Based on the results shown in Table 5.1, we can deduce that *male_adult_white* (the favored category) has the highest probability of being classified as class 1 (favorable class). This is not surprising as the category has the highest number of samples, and also, the particular category forms the baseline of our sampling algorithm because it is somehow favored among all other categories. This shows that feeding this type of data into any algorithm without any biased tackling method will result in a bias prediction, as we see in the results in Table 4.1, which indicates the bias of the no-sampling method.

Based on the results in Table 5.2, we can see an increase in all the protected attribute class 1 (favorable class), but one other thing to note is the substantial increase in the protected classes for the favored protected attribute (*male_adult_white*) while only a small increase in another group. This was later reflected in the results of proportional sampling of

Table 5.2: Proportional Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Adult Income Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.760	0.240	0.08
Female_Adult_White	0.636	0.364	0.10
Female_Young_Other	0.958	0.042	0.01
Female_Young_White	0.961	0.039	0.01
Male_Adult_Other	0.464	0.536	0.15
Male_Adult_White	0.367	0.633	0.13
Male_Young_Other	0.958	0.042	0.01
Male_Young_White	0.932	0.068	0.02

the adult income dataset in Table 4.1, especially a far worse EOD of 0.66. This shows that an increase in positive class for favorable protected groups can have a detrimental effect on an algorithm's prediction and fairness.

Based on the results in Table 5.3, we see a slight increase in class 1 of all the protected categories except for the favorable class. This increment is too small to make any decisive shift in both prediction and fairness, and this contributes to why the oversampling method did not yield a better EOD in comparison to the baseline. Also, the increment in the favored group counteracts the effect of the increment in the unfavored group, thereby leading to a more pronounced bias in the prediction of this algorithm. Also, the KL divergence shows the difference in the distribution of the no-sampling method, and oversampling is negligible across all the protected categories, which shows why the sampling method struggles in both prediction and fairness.

Table 5.3: Oversample Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Adult Income Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.910	0.090	0.000
Female_Adult_White	0.847	0.153	0.000
Female_Young_Other	0.985	0.015	0.000
Female_Young_White	0.988	0.012	0.000
Male_Adult_Other	0.741	0259	0.000
Male_Adult_White	0.625	0.375	0.000
Male_Young_Other	0.982	0.018	0.001
Male_Young_White	0.979	0.021	0.000

Table 5.4: PC-SMOTE Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Adult Income Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.686	0.314	0.15
Female_Adult_White	0.731	0.269	0.03
Female_Young_Other	0.708	0.292	0.28
Female_Young_White	0.669	0.331	0.35
Male_Adult_Other	0.671	0.329	0.11
Male_Adult_White	0.625	0.375	0.00
Male_Young_Other	0.669	0.331	0.34
Male_Young_White	0.721	0.279	0.23

Table 5.5: PC-ADASYN Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Adult Income Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.914	0.086	0.00
Female_Adult_White	0.842	0.158	0.00
Female_Young_Other	0.989	0.011	0.00
Female_Young_White	0.984	0.016	0.00
Male_Adult_Other	0.740	0.260	0.00
Male_Adult_White	0.625	0.375	0.00
Male_Young_Other	0.984	0.016	0.00
Male_Young_White	0.977	0.023	0.00

Based on the results in Table 5.4, we can see a significant increase in the favorable class (class 1) of all the protected attributes except the *male_adult_white* This corroborates with the performance of the algorithm as we can see a substantial increase in EOD of PC-SMOTE in Table 4.1. One important thing to note in these results is the substantial increase in the unfavorable group such as $female\_adult\_other$; this is necessary to improve the fairness of the algorithm. This shows the interpolation strategy used in this sampling method is effective enough to move samples from a region of unfairness to a fairness region, thereby mitigating bias.

Based on the results in Table 5.5, we can see the sampling method has only a negligible increase in all the other protected categories. This contradicts what we saw in Table 5.4, where there was a substantial improvement. This can be attributed to two reasons. The first reason is that the underlying mechanisms of our proposed PC-ADASYN do not need to

increase the class categories to improve fairness. The reason is that the relationship between our sampling methods and class labels needs to be studied extensively to better show whether we need an increase in favorable class probability to improve fairness.

Overall, we can see different performances in the sampling methods regarding the changing distribution of each class in comparison to the sampling methods. Each sampling method improves the probability of a favorable class, but some are negligible.

CHAPTER SIX

CONCLUSIONS

6.1 Contributions

The broad problem of fairness in machine learning is significant in that the prevalence of AI and ML systems today is having a major impact on people's lives and livelihoods. While attention to fair ML has increased substantially, there continues to be a need for methods to advance fair ML without negatively impacting ML performance. Based on an in-depth review of the literature and the above need for this type of work, the methods reported here make the following contributions:

1. The commonly-held assumption that there exists an inherent tradeoff between fairness and performance (i.e., accuracy) in machine learning is challenged with evidence provided to support this challenge. In particular, the results in this paper indicate that such a tradeoff can be mitigated, suggesting that any tradeoff is most likely tied to how the data is being managed.

2. Four novel pre-processing methods for sampling data are presented based on applying a multi-category sampling strategy using data captured in protected categories. The methods proceed from the assumption and corresponding hypothesis that balancing the data based on these multi-category properties can increase fairness without adversely affecting machine learning model performance.

3. Experimental results are presented using five data sets studied extensively within the fair ML community. The experiments include comparisons with traditional methods of training with no pre-processing to demonstrate the relative effects of the proposed

methods. The results demonstrate that two of the proposed methods, the Protected-Category Synthetic Minority Oversampling Technique (PC-SMOTE) and Protected-Category Adaptive Synthetic sampling (PC-ADASYN), are particularly effective in improving both fairness and performance.

4. A detailed analysis relating the potential effects of underfitting and overfitting on fairness is presented by examining different levels in a decision tree model, with and without using the proposed sampling methods. The results demonstrate the ability of the proposed methods to identify an ideal level of the tree where both fairness and accuracy are maximized.

5. A detailed analysis of the impact of our sampling method on the distributions of class labels was performed. We examined the conditional probabilities of each class in each protected attribute for both no-sampling and our sampling method and compared the KL divergence between the no-sampling method and our sampling methods to deduce if there is any difference between these distributions.

As a result of the above contributions, this work represents a significant step forward in addressing concerns of fairness in machine learning. A key takeaway from the methods and results reported here is that fairness can be addressed without compromising model performance. As a result, the core results of this thesis has been published in [78]

6.2 Summary

In this thesis, the issue of bias in ML predictions was investigated, and a method was developed based on combining protected variables into a new multi-category. In particular, the focus was on the question that has been suggested in the literature of a bias-performance tradeoff and seek a method to mitigate this tradeoff. The proposed new multi-category approach reflects the multifaceted identity of individuals, acknowledging the complex

interplay of attributes that define real-world scenarios. Given the inherent imbalance in this multi-category, four sampling methods tailored to these complex categorizations, rather than traditional class labels, were developed. For purposes of applying a baseline classifier, decision trees were trained, and the effectiveness of these methods was evaluated using five datasets that are often employed in fairness studies. The performance of the methods was compared against baseline methods of no sampling using accuracy, macro F1, Equalized Odds Difference (EOD), Statistical Parity (SP), Disparate Impact (DI), and Prediction Equality (PE) as the evaluation metrics.

The results of the experiments indicate that two of the newly developed sampling techniques—PC-SMOTE and PC-ADASYN—successfully enhance fairness without compromising accuracy. Remarkably, in some cases, these methods also improved accuracy, thus providing evidence counter to the popular claims of a fairness-performance tradeoff. Further analysis of the impact of maximum tree depth on model performance revealed that, while increasing depth initially boosts accuracy, it eventually leads to a decline. Conversely, increasing depth adversely affects fairness, highlighting the challenge of balancing complexity with equity. However, optimal tree depths were identified that simultaneously enhance accuracy and EOD, underscoring the possibility of achieving equity without sacrificing performance.

Ultimately, these findings challenge prevailing notions of an implicit performance-fairness tradeoff within bias mitigation research, suggesting that carefully designed bias mitigation strategies have the ability to sidestep this trade-off. Our approach sets a new precedent for developing more equitable predictive algorithms by redefining how protected attributes are utilized in model training.

6.3 Limitations

The very nature of this study is such that it is not possible to address all of the issues surrounding fairness and the so-called fairness-performance tradeoff. As such, there exist limitations in the work reported here. Even so, it is our hope and intent for this work to suggest additional avenues of exploration in this important area.

One limitation of this study is that our sampling method was not specifically designed to optimize for arbitrary fairness metrics. Stated another way, since inherent tradeoffs often exist between the available set of fairness metrics, the decision was made to focus on an approach that was metric agnostic, recognizing that the results could have differed for other metrics. This is also part of the reason why we saw different behaviors between EOD, SP, DI, and PE.

6.4 Future Works

In addition, it is acknowledged that, while the underlying ML method should not be relevant to the method proposed, this has not actually been tested. Therefore, in the future, this research will be extended by considering the impacts of other ML algorithms such as logistic regression, fuzzy ID3, K-nearest neighbor, neural network, and ensemble methods such as random forests or gradient-boosted trees to assess the generalizability of our new sampling methods. The purpose of such a study would be to verify that our methods are independent of the ML algorithm employed. Furthermore, this would help validate whether the observed improvements in fairness and accuracy are model-specific or can be applied universally.

Additionally, it is acknowledged that only five distinct data sets were considered—data sets that have been studied extensively in the field. This raises a concern that methods are being tailored to these data rather than addressing the broader issue of fairness in ML. To

address this, experiments with larger and more diverse datasets are planned to provide deeper insights into the scalability and robustness of our techniques. Another area for future work is to refine our multi-category sampling approach by incorporating more granular subdivisions of protected categories, potentially revealing subtler biases and providing a more nuanced understanding of fairness.

Finally, it is recognized that alternative methods have been proposed for bias mitigation, and these methods have not been studied in this work at all. Examples include FairSMOTE, CounterFactual Fairness, etc. Future work would include comparisons to more sampling strategies. More direct comparison of the proposed methods with in-processing and post-processing methods will be done. For example, incorporating in-processing methods, such as regularization [54], or a post-processing method such as the Randomized Threshold Optimizer [5] will be explored as possible means to obtain further improvements in both fairness and performance.

One of the future endeavors we plan to embark on is optimization for multiple fairness. This can be in the form of multi-objective optimization or a selection of different methods that optimize different metrics and combine them together.

REFERENCES CITED

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69, 2018.

[2] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1418–1426, 2019.

[3] Astha Agrawal, Herna L Viktor, and Eric Paquet. SCUT: Multi-class imbalanced data classification using smote and cluster-based undersampling. In *IEEE International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3k)*, volume 1, pages 226–234, 2015.

[4] Gulnaz Ahmed, Meng Joo Er, Mian Muhammad Sadiq Fareed, Shahid Zikria, Saqib Mahmood, Jiao He, Muhammad Asad, Syeda Fizzah Jilani, and Muhammad Aslam. DAD-Net: Classification of alzheimer's disease using ADASYN oversampling technique and optimized neural network. *Molecules*, 27(20):7085, 2022.

[5] Ibrahim Alabdulmohsin and Mario Lucic. A near-optimal algorithm for debiasing trained machine learning models. In *35th Conference on Neural Information Processing Systems*, 2021.

[6] Yuan Bao and Sibo Yang. Two novel smote methods for solving imbalanced classification problems. *IEEE Access*, 11:5816–5823, 2023.

[7] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[8] John J Benedetto and Paulo JSG Ferreira. *Modern sampling theory: mathematics and applications*. Springer Science & Business Media, 2012.

[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[10] Dheeraj Bhaskaruni, Hui Hu, and Chao Lan. Improving prediction fairness via model ensemble. In *IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1810–1814, 2019.

[11] Rok Blagus and Lara Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14:1–16, 2013.

[12] B Bouslah, A Gharbi, and Robert Pellerin. Integrated production, sampling quality control and maintenance of deteriorating production systems with AOQL constraint. *Omega*, 61:110–126, 2016.

[13] Max Bramer. Avoiding overfitting of decision trees. *Principles of data mining*, pages 119–134, 2007.

[14] Jakob Brandt and Emil Lanzén. A comparative review of SMOTE and ADASYN in imbalanced data classification. *Digitala Vetenskapliga Arkivet*, 2021.

[15] Dick J Brus and Gerard BM Heuvelink. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138(1-2):86–95, 2007.

[16] Andrew Burt. How to fight discrimination in AI. `https://hbr.org/2020/08/how-to-fight-discrimination-in-ai`, 2020. Harvard Business Review, Accessed 07/12/2024.

[17] Toon Calders and Indrè Žliobaitè. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pages 23–33. Springer, 2013.

[18] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in neural information processing systems*, 2017.

[19] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56:1–38, 2023.

[20] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishno. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[21] Junyi Chai and Xiaoqian Wang. Fairness with adaptive weights. In *International Conference on Machine Learning*, pages 2853–2866. PMLR, 2022.

[22] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[23] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM transactions on software engineering and methodology*, 32:1–30, 2023.

[24] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019.

[25] Umer Daraz, Jinbiao Wu, Mohammed Ahmed Alomair, and Luai Abdulla Aldoghan. New classes of difference cum-ratio-type exponential estimators for a finite population variance in stratified random sampling. *Heliyon*, 10(13), 2024.

[26] Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D Castillo, and Rama Chellappa. PASS: Protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15087–15096, 2021.

[27] William P Dommermuth. *The use of sampling in marketing research*. Marketing Classics Press, 2011.

[28] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[29] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[30] Martin Ebers, Veronica R. S. Hoch, Frank Rosenkranz, Hannah Ruschemeier, and Björn Steinrötter. The European Commission's proposal for an artificial intelligence act–a critical assessment by members of the robotics and AI law society (RAILS). *MDPI*, 4(4):589–603, 2021.

[31] Rokaya Eltehewy, Ahmed Abouelfarag, and Sherine Nagy Saleh. Efficient classification of imbalanced natural disasters data using generative adversarial networks for data augmentation. *ISPRS International Journal of Geo-Information*, 12(6):245, 2023.

[32] Robert Wall Emerson. Convenience sampling, random sampling, and snowball sampling: How does sampling affect the validity of research? *Journal of visual impairment & blindness*, 109(2):164–168, 2015.

[33] Tobias Fahse, Viktoria Huber, and Benjamin van Giffen. Managing bias in machine learning projects. In *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*, pages 94–109. Springer, 2021.

[34] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[35] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the SIAM international conference on data mining*, pages 144–152, 2016.

[36] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64:136–143, 2021.

[37] Jesús F.Salgado, Silvia Moscoso, Antonio L. García-Izquierdo, and Neil R. Anderson. *Shaping inclusive workplaces through social dialogue*. Springer, 2017.

[38] Vicente García, José Salvador Sánchez, and Ramón Alberto Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2012.

[39] Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, pages 1445–1459, 2012.

[40] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 2016.

[41] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.

[42] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[43] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[44] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 2023.

[45] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 994–1006, 2012.

[46] Qiujun Huang, Jingli Mao, and Yong Liu. An improved grid search algorithm of svr parameters optimization. In *IEEE 14th International Conference on Communication Technology*, pages 1022–1026, 2012.

[47] Brian HuZhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[48] Rufai Iliyasu and Ilker Etikan. Comparison of quota sampling and stratified random sampling. *Biometrics Biostatistics International Journal*, 10(1):24–27, 2021.

[49] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. FAR: A fairness-aware ensemble framework. In *IEEE international conference on big data (Big Data)*, pages 1375–1380, 2019.

[50] Patrick Janssen and Bert M. Sadowski. Bias in algorithms: On the trade-off between accuracy and fairness. In *23rd Biennial Conference of the International Telecommunications Society*, 2021.

[51] Nicole Blair Johnson, Locola D Hayes, Kathryn Brown, Elizabeth C Hoo, Kathleen A Ethier, Centers for Disease Control, and Prevention (CDC). CDC national health report: Leading causes of morbidity and mortality and associated behavioral risk and protective factors–United States, 2005-2013. *MMWR*, 63(4):3–27, 2014.

[52] Firuz Kamalov and Dmitry Denisov. Gamma distribution-based sampling for imbalanced data. *Knowledge-Based Systems*, 207:106368, 2020.

[53] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33:1–33, 2012.

[54] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, pages 35–50, September 2012.

[55] Marzieh Karimi-Haghighi and Carlos Castillo. Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 210–214, 2017.

[56] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. https://archive.ics.uci.edu, 2024.

[57] Nima Kordzadeh and Maryam Ghasemaghaei. Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409, 2022.

[58] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the World Wide Web Conference*, pages 853–862, 2018.

[59] Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. Mitigating gender bias in machine learning data sets. In *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020*, pages 12–26, 2020.

[60] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[61] Jie Liu. Importance-SMOTE: A synthetic minority oversampling method for noisy imbalanced data. *Soft Computing*, 26(3):1141–1163, 2022.

[62] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

[63] Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. *Transactions on Machine Learning Research*, 2022.

[64] Ivy W. Maina, Tanisha D. Belton, Sara Ginzberg, Ajit Singh, and Tiffani J. Johnson. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social science and medicine*, 199:219–229, 2018.

[65] Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Said Hacid, and Hassan Zeineddine. An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, 7:93010–93022, 2019.

[66] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *11th International Conference on Information and Communication Systems (ICICS)*, pages 243–248. IEEE, 2020.

[67] Sayed A Mostafa and Ibrahim A Ahmad. Recent developments in systematic sampling: A review. *Journal of Statistical Theory and Practice*, 12(2):290–310, 2018.

[68] Nhlakanipho Mqadi, Nalindren Naicker, and Timothy Adeliyi. A SMOTE based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection. *International Journal of Computing and Digital Systems*, 10(1):277–286, 2021.

[69] Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics: Methodology and distribution*, pages 123–150. Springer, 1992.

[70] Shagofah Noor, Omid Tajik, and Jawad Golzar. Simple random sampling. *International Journal of Education & Language Studies*, 1(2):78–82, 2022.

[71] R Patel Brijain and Kaushik K Rana. A survey on decision tree algorithm for classification. *International journal of Engineering development and research*, 2(1):1–5, 2014.

[72] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[73] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. FairMask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering*, 49:2426–2439, 2022.

[74] Adam Perzynski, Kristen A. Berg, Charles Thomas, Anupama Cemballi, Tristan Smith, Sarah Shick, Douglas Gunzler, and Ashwini R. Sehgal. Racial discrimination and economic factors in redlining of Ohio neighborhoods. *Du Bois Review: Social Science Research on Race*, 20:293–309, 2023.

[75] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55, 2022.

[76] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in neural information processing systems*, 2017.

[77] Gideon Popoola and Kayode Oyeniran. Faco: A novel hybrid feature selection algorithm for high-dimensional data classification. In *SoutheastCon 2024*, pages 61–68. IEEE, 2024.

[78] Gideon Popoola and John Sheppard. Investigating and mitigating the performance–fairness tradeoff via protected-category sampling. *Electronics*, 13(15):3024, 2024.

[79] Sai Prasad Potharaju and M Sreedevi. An improved prediction of kidney disease using smote. *Indian Journal of Science and Technology*, 9(31):1–7, 2016.

[80] Sanja Rančić, Sandro Radovanović, and Boris Delibašić. Investigating oversampling techniques for fair machine learning models. In *Decision Support Systems XI: Decision Support Systems, Analytics and Technologies in Response to Global Crisis Management: 7th International Conference on Decision Support System Technology, ICDSST 2021, Loughborough, UK, May 26–28, 2021, Proceedings*, pages 110–123. Springer, 2021.

[81] Sanja Rančić, Sandro Radovanović, and Boris Delibašić. Investigating oversampling techniques for fair machine learning models. In *Decision Support Systems XI: Decision Support Systems, Analytics and Technologies in Response to Global Crisis Management: 7th International Conference on Decision Support System Technology, 2021, Proceedings*, pages 110–123. Springer, 2021.

[82] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. In *Advances in neural information processing systems*, pages 361–371, 2020.

[83] Bengt Rosén. On sampling with probability proportional to size. *Journal of statistical planning and inference*, 62(2):159–191, 1997.

[84] Kaylee Rosenberger, Emily Schumacher, Alissa Brown, and Sean Hoban. Proportional sampling strategy often captures more genetic diversity when population sizes vary. *Biological Conservation*, 261:109261, 2021.

[85] Salvatore Ruggieri. Using t-closeness anonymity to control for non-discrimination. *Transaction of Data Privacy*, 2:99–129, 2014.

[86] Teresa Salazar, Miriam Seoane Santos, Helder Araújo, and Pedro Henriques Abreu. FAWOS: Fairness-aware oversampling algorithm based on distributions of sensitive attributes. *IEEE Access*, 9:81370–81379, 2021.

[87] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[88] Philip Sedgwick. Convenience sampling. *British Medical Journal*, 347, 2013.

[89] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. Representation bias in data: a survey on identification and resolution techniques. *ACM Computing Surveys*, 55:1–39, 2023.

[90] Kyarash Shahriari and Mana Shahriari. IEEE standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 197–201, 2017.

[91] Gaganpreet Sharma. Pros and cons of different sampling techniques. *International journal of applied research*, 3(7):749–752, 2017.

[92] Swati Shilaskar and Ashok Ghatol. Diagnosis system for imbalanced multi-minority medical dataset. *Soft Computing*, 23(13):4789–4799, 2019.

[93] Alison Siegler and William Admussen. Discovering racial discrimination by the police. *Northwestern University Law Review*, 115:987–1054, 2021.

[94] Louis-Philippe Sondeck, Maryline Laurent, and Vincent Frey. The semantic discrimination rate metric for privacy measurements which questions the benefit of t-closeness over l-diversity. In *14th International Conference on Security and Cryptography*, volume 6, pages 285–294, 2017.

[95] Ryosuke Sonoda. Fair oversampling technique using heterogeneous clusters. *Information Sciences*, 640:119059, 2023.

[96] Justin P. Steil, Len Albright, Jacob S. Rugh, and Douglas S. Massey. The social structure of mortgage discrimination. *Housing studies*, 33:759–776, 2018.

[97] Lee-Jen Wu Suen, Hui-Man Huang, and Hao-Hsien Lee. A comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, 61(3):105, 2014.

[98] Bo Tang and Haibo He. Kerneladasyn: Kernel based adaptive synthetic data generation for imbalanced learning. In *2015 IEEE congress on evolutionary computation (CEC)*, pages 664–671. IEEE, 2015.

[99] Suryakanthi Tangirala. Evaluating the impact of Gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2):612–619, 2020.

[100] TobiasFahse, Viktoria Huber, and Benjamin van Giffen. Managing bias in machine learning projects. In *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*, pages 94–109. Springer, 2021.

[101] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

[102] Qing Ren Wang and Ching Y Suen. Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):406–417, 1984.

[103] Yun Xu and Royston Goodacre. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, 2(3):249–262, 2018.

[104] Shen Yan, Hsien te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2010.

[105] I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

[106] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[107] Masoumeh Zareapoor, Pourya Shamsolmoali, and Jie Yang. Oversampling adversarial network for class-imbalanced fault diagnosis. *Mechanical Systems and Signal Processing*, 149:107175, 2021.

[108] Zhe Zhang and Daniel B. Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2017.

[109] Zhuoyuan Zheng, Yunpeng Cai, and Ye Li. Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5):1017–1037, 2015.

APPENDICES

APPENDIX A

MORE EXPERIMENTAL TREE RESULTS

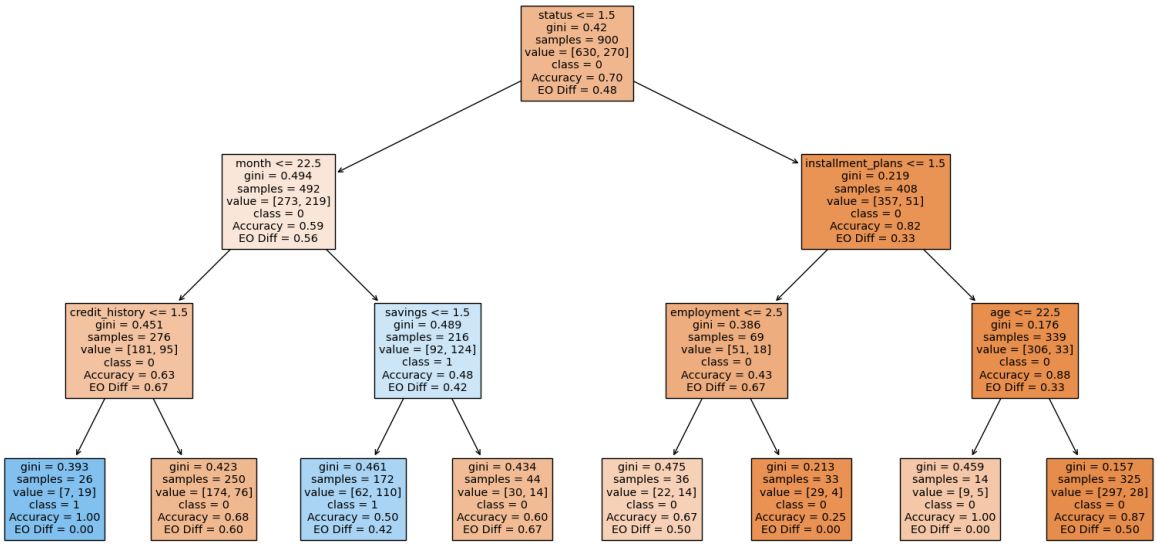

Figure A.1: Example decision tree trained on German Credit with no sampling

Below are more results from our experiments. The first results show the tree from no-sampling and ADASYN from other datasets.

A.1 German Credit Dataset

These figures are the decision tree generated for the German Credit dataset. The A.1 shows the no-sampling decision tree while the A.2 shows the decision tree for the ADASYN sampling method.

A.2 COMPAS Dataset

These figures are the decision tree generated for the COMPASt dataset. The A.3 shows the no-sampling decision tree while the A.4 shows the decision tree for the ADASYN sampling method.

A.3 Bank Credit Default

These figures are the decision tree generated for the Bank Credit Default dataset. The A.5 shows the no-sampling decision tree while the A.6 shows the decision tree for the ADASYN sampling method.

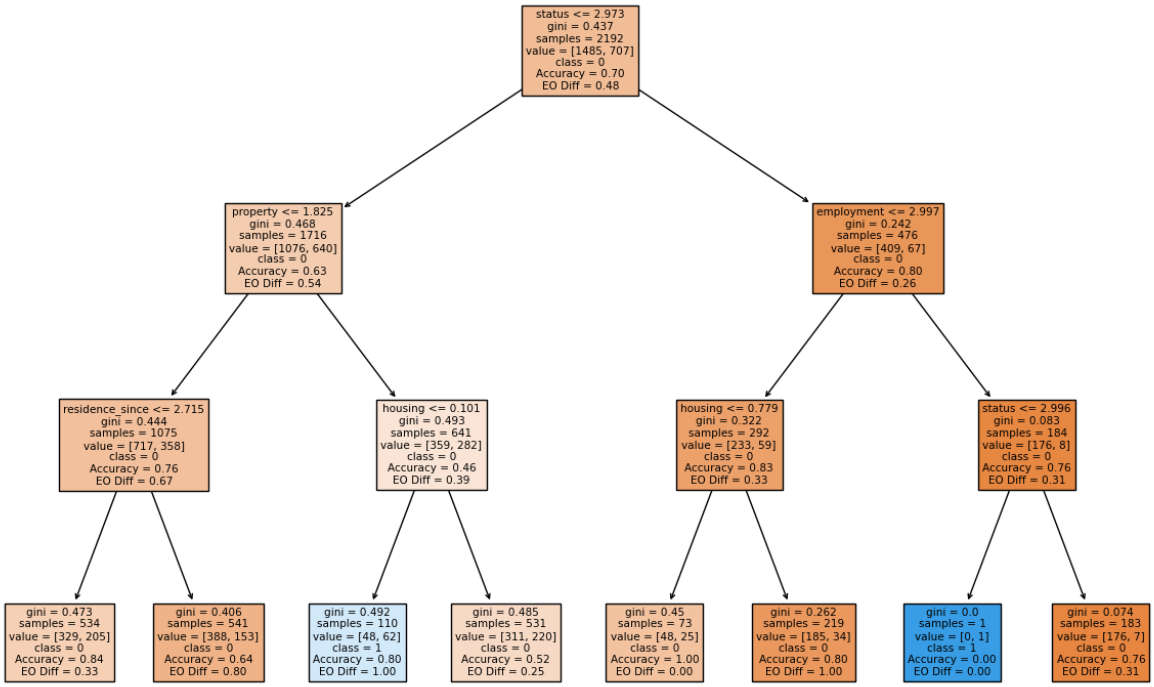

Figure A.2: Decision tree of PC-ADASYN on German Credit

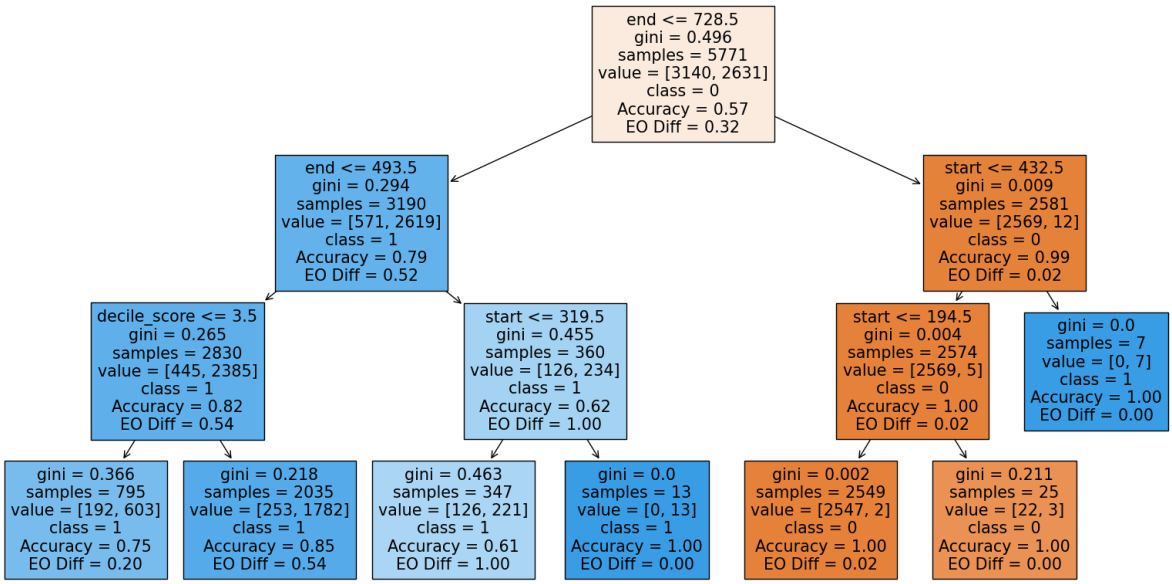

Figure A.3: Example decision tree trained on COMPAS with no sampling

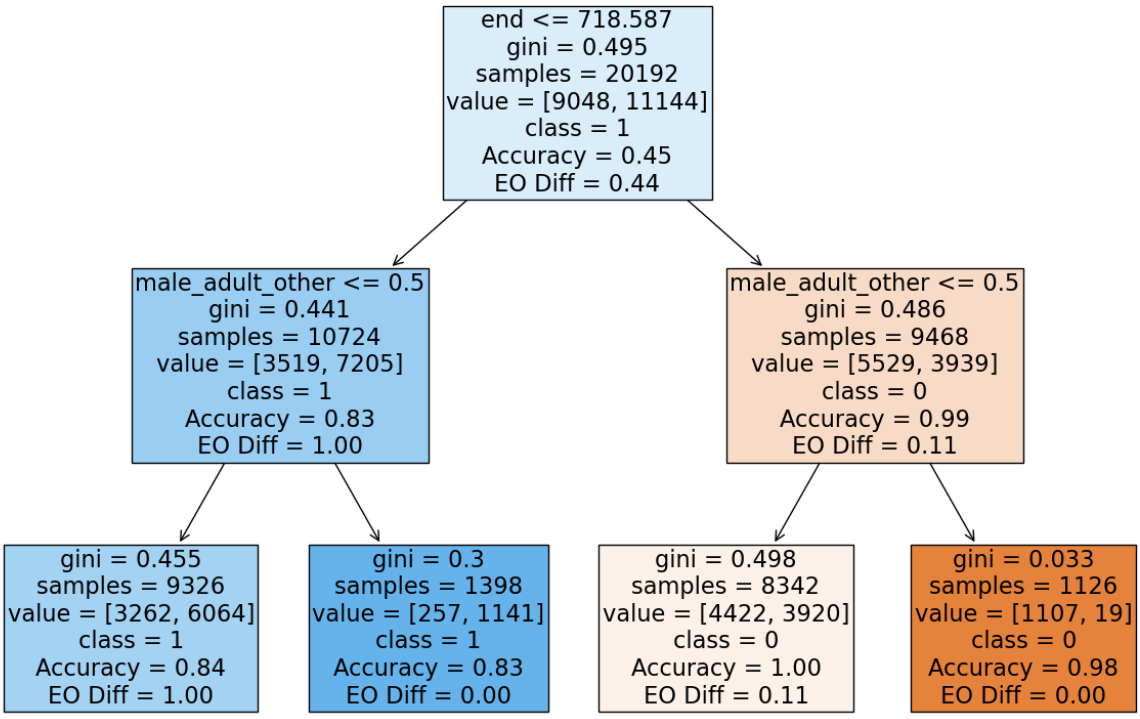

Figure A.4: Decision tree of PC-ADASYN on COMPAS

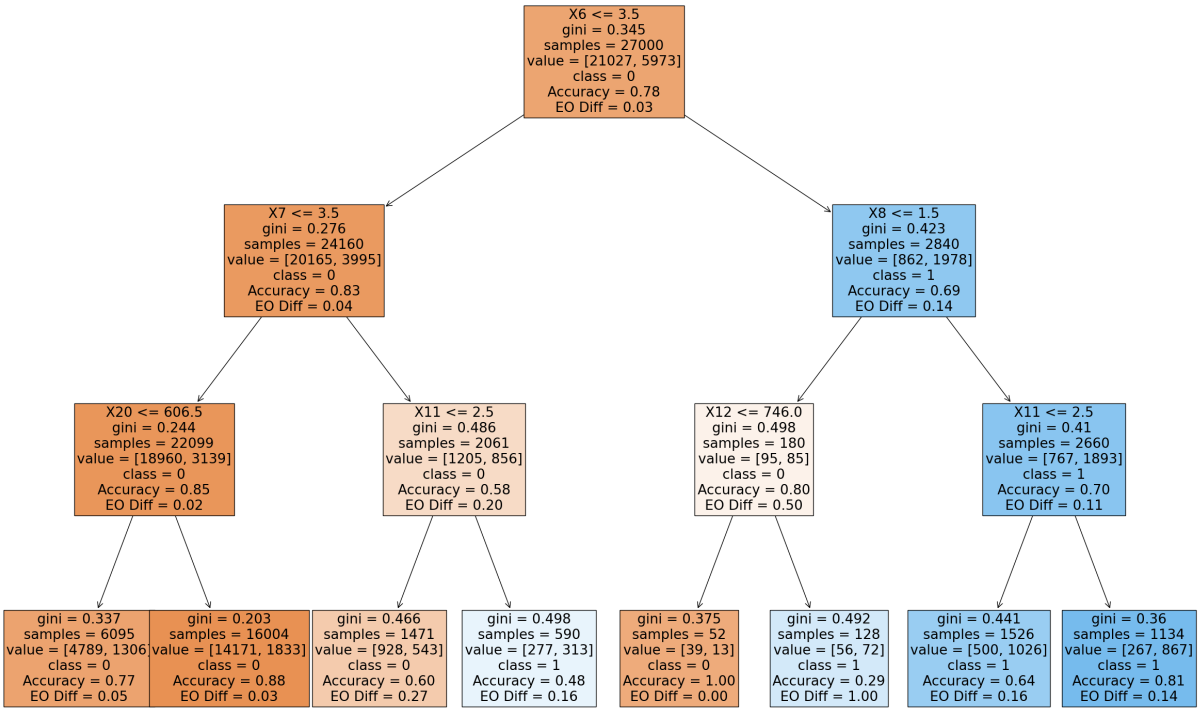

Figure A.5: Example decision tree trained on Bank Credit Default with no sampling

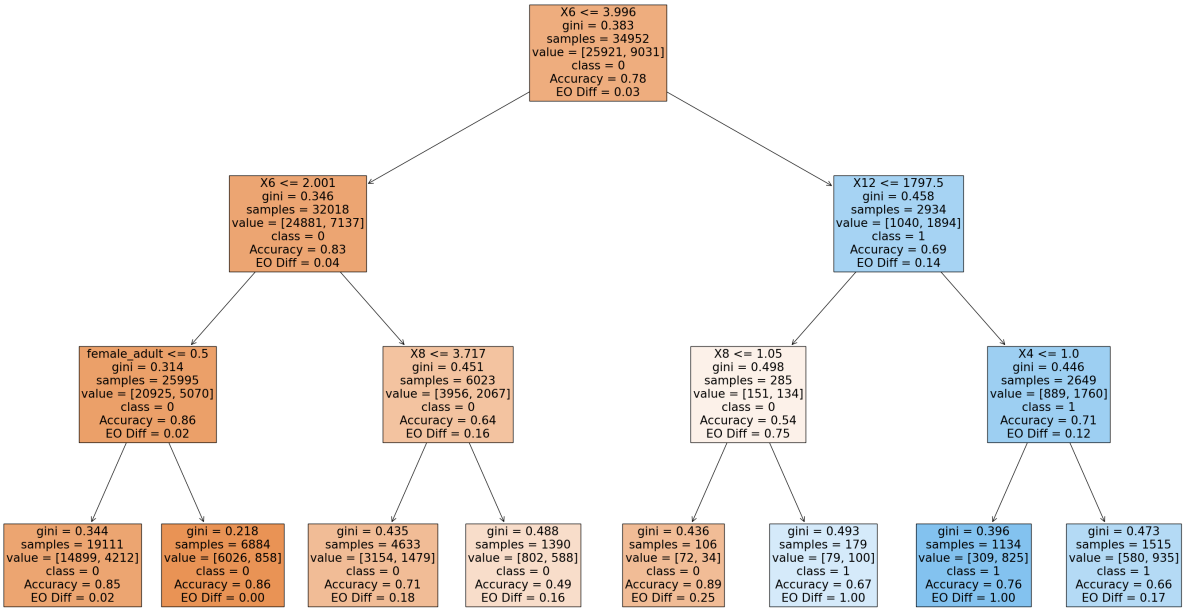

Figure A.6: Decision tree of PC-ADASYN on Bank Credit Default

A.4 CDC Diabetes Dataset

These figures are the decision tree generated for the CDC Diabetes dataset. The A.7 shows the no-sampling decision tree while the A.8 shows the decision tree for the ADASYN sampling method.

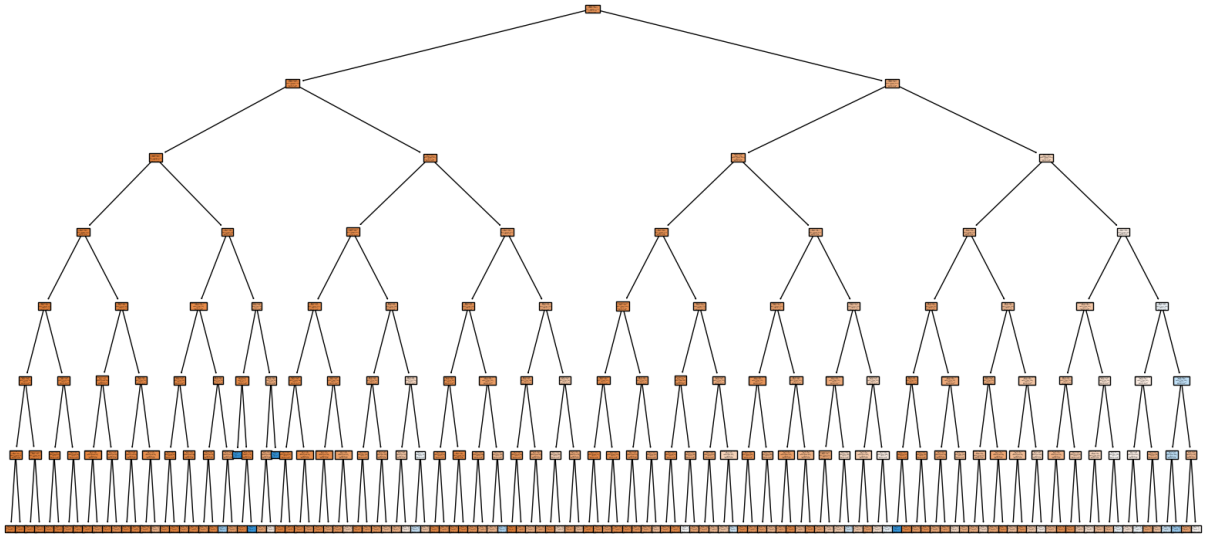

Figure A.7: Example decision tree trained on CDC Diabetes with no sampling

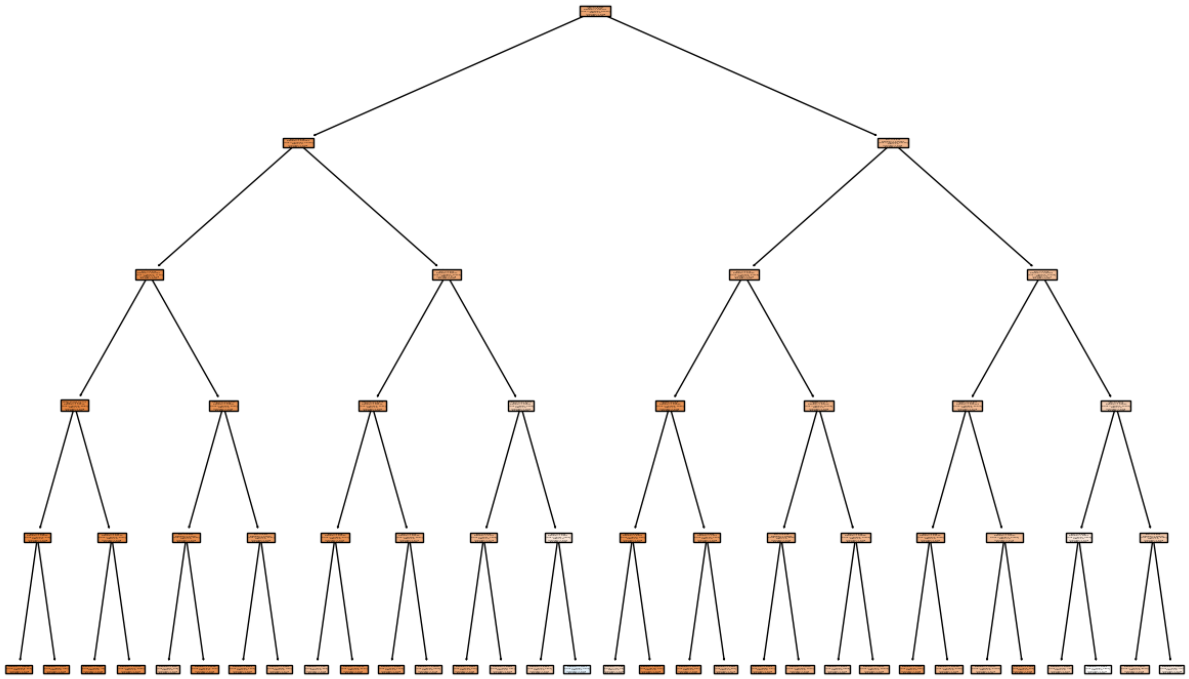

Figure A.8: Decision tree of PC-ADASYN on CDC Diabetes

APPENDIX B

KL DIVERGENCE RESULTS

Table B.1: No-Sampling Conditional Probabilities of each class and each protected attribute of German Credit Dataset

Protected_Category	Class_0	Class_1
Female_Adult	0.674	0.326
Female_Young	0.567	0.433
Male_Adult	0.747	0.253
Male_Young	0.587	0.413

Table B.2: Oversample Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for German Credit Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.738	0.262	0.010
Female_Young	0.614	0.386	0.004
Male_Adult	0.659	0.340	0.017
Male_Young	0.496	0.504	0.016

B.1 German Credit

The tables are the KL divergence table generated for the German Credit dataset. The B.1 shows the no-sampling probability distribution for each class, table B.2 shows the oversampling probability distribution and its KL divergence in comparison to the no-sampling, table B.3 shows the proportional sampling probability distribution for each class and its KL divergence in comparison to the no-sampling method, table B.4 shows the PC-SMOTE probability distribution and its KL divergence in comparison to the no-sampling method, and table B.5 show the PC-SMOTE probability distribution and it KL divergence in comparison to the no-sampling method.

B.2 COMPAS Dataset

The tables are the KL divergence table generated for the COMPAS dataset. The B.6 shows the no-sampling probability distribution for each class, table B.7 shows the oversampling probability distribution and its KL divergence in comparison to the no-sampling, table B.8 shows the proportional sampling probability distribution for each class and its KL divergence in comparison to the no-sampling method, table B.9 shows the PC-SMOTE probability distribution and its KL divergence in comparison to the no-sampling method, and table B.10 show the PC-SMOTE probability distribution and it KL divergence in comparison to the no-sampling method.

Table B.3: Proportional Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for German Credit Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.542	0.458	0.036
Female_Young	0.306	0.694	0.145
Male_Adult	0.54	0.46	0.090
Male_Young	0.456	0.544	0.034

Table B.4: PC-SMOTE Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for German Credit Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.666	0.334	0.000
Female_Young	0.646	0.354	0.013
Male_Adult	0.746	0.254	0.000
Male_Young	0.651	0.349	0.009

Table B.5: PC-ADASYN Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for German Credit Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.666	0.334	0.000
Female_Young	0.646	0.354	0.013
Male_Adult	0.746	0.254	0.000
Male_Young	0.651	0.349	0.009

Table B.6: No-Sampling Conditional Probabilities of each class and each protected attribute of COMPAS Dataset

Protected_category	Class_0	Class_1
Female_Adult_Other	0.658	0.342
Female_Adult_White	0.631	0.369
Female_Young_Other	0.599	0.401
Female_Young_White	0.662	0.337
Male_Adult_Other	0.533	0.464
Male_Adult_White	0.613	0.387
Male_Young_Other	0.373	0.626
Male_Young_White	0.489	0.511

Table B.7: Oversample Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for COMPAS Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.638	0.362	0.000
Female_Adult_White	0.600	0.400	0.002
Female_Young_Other	0.607	0.393	0.000
Female_Young_White	0.683	0.317	0.000
Male_Adult_Other	0.539	0.461	0.000
Male_Adult_White	0.625	0.375	0.000
Male_Young_Other	0.414	0.586	0.003
Male_Young_White	0.443	0.557	0.004

Table B.8: Proportional Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for COMPAS Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.627	0.373	0.002
Female_Adult_White	0.575	0.425	0.006
Female_Young_Other	0.566	0.434	0.002
Female_Young_White	0.591	0.409	0.010
Male_Adult_Other	0.502	0.498	0.002
Male_Adult_White	0.557	0.443	0.006
Male_Young_Other	0.350	0.650	0.001
Male_Young_White	0.394	0.606	0.018

Table B.9: PC-SMOTE Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for COMPAS Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.456	0.544	0.082
Female_Adult_White	0.437	0.563	0.075
Female_Young_Other	0.410	0.599	0.071
Female_Young_White	0.414	0.586	0.124
Male_Adult_Other	0.540	0.460	0.000
Male_Adult_White	0.518	0.482	0.018
Male_Young_Other	0.398	0601	0.001
Male_Young_White	0.410	0.599	0.012

Table B.10: PC-ADASYN Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for COMPAS Dataset

Protected_category	Class_0	Class_1	KL_divergence
Female_Adult_Other	0.456	0.544	0.082
Female_Adult_White	0.437	0.563	0.075
Female_Young_Other	0.410	0.599	0.071
Female_Young_White	0.414	0.586	0.124
Male_Adult_Other	0.540	0.460	0.000
Male_Adult_White	0.518	0.482	0.018
Male_Young_Other	0.398	0601	0.001
Male_Young_White	0.410	0.599	0.012

Table B.11: No-Sampling Conditional Probabilities of each class and each protected attribute of Bank Credit Default Dataset

Protected_Category	Class_0	Class_1
Female_Adult	0.799	0.201
Female_Young	0.783	0.217
Male_Adult	0.756	0.244
Male_Young	0.760	0.240

B.3 Bank Credit Default Dataset

The tables are the KL divergence table generated for the Bank Credit Default dataset. The B.11 shows the no-sampling probability distribution for each class, table B.12 shows the oversampling probability distribution and its KL divergence in comparison to the no-sampling, table B.13 shows the proportional sampling probability distribution for each class and its KL divergence in comparison to the no-sampling method, table B.14 shows the PC-SMOTE probability distribution and its KL divergence in comparison to the no-sampling method, and table B.15 show the PC-SMOTE probability distribution and it KL divergence in comparison to the no-sampling method.

B.4 CDC Diabetes Dataset

The tables are the KL divergence table generated for the CDC Diabetes dataset. The B.16 shows the no-sampling probability distribution for each class, table B.17 shows the oversampling probability distribution and its KL divergence in comparison to the no-

Table B.12: Oversample Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Bank Credit Default Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.799	0.201	0.000
Female_Young	0.783	0.217	0.000
Male_Adult	0.756	0.244	0.000
Male_Young	0.760	0.240	0.000

sampling, table B.18 shows the proportional sampling probability distribution for each class and its KL divergence in comparison to the no-sampling method, table B.19 shows the PC-SMOTE probability distribution and its KL divergence in comparison to the no-sampling method, and table B.20 show the PC-SMOTE probability distribution and it KL divergence in comparison to the no-sampling method.

Table B.13: Proportional Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Bank Credit Default Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.528	0.472	0.159
Female_Young	0.505	0.495	0.164
Male_Adult	0.468	0.532	0.171
Male_Young	0.482	0.518	0.162

Table B.14: PC-SMOTE Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Adult Income Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.799	0.201	0.000
Female_Young	0.741	0.259	0.005
Male_Adult	0.730	0.270	0.001
Male_Young	0.695	0.305	0.105

Table B.15: PC-ADASYN Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for Adult Income Dataset

Protected_Category	Class_0	Class_1	KL_Divergnce
Female_Adult	0.799	0.201	0.000
Female_Young	0.741	0.259	0.005
Male_Adult	0.730	0.270	0.001
Male_Young	0.695	0.305	0.105

Table B.16: No-Sampling Conditional Probabilities of each class and each protected attribute of CDC Diabetes Dataset

Protected_Category	Class_0	Class_1
Female_Adult	0.815	0.185
Female_Young	0.887	0.113
Male_Adult	0.763	0.237
Male_Young	0.871	0.129

Table B.17: Oversample Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for CDC Diabetes Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.760	0.240	0.008
Female_Young	0.872	0.127	0.000
Male_Adult	0.813	0.187	0.007
Male_Young	0.887	0.113	0.001

Table B.18: Proportional Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for CDC Diabetes Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.414	0.587	0.338
Female_Young	0.561	0.439	0.253
Male_Adult	0.342	0.658	0.369
Male_Young	0.523	0.477	0.275

Table B.19: PC-SMOTE Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for CDC Diabetes Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.707	0.293	0.030
Female_Young	0.887	0.113	0.000
Male_Adult	0.682	0.318	0.157
Male_Young	0.831	0.169	0.006

Table B.20: PC-ADASYN Sampling method's Conditional Probability for Each Class and Each Protected with KL Divergence for CDC Diabetes Dataset

Protected_Category	Class_0	Class_1	KL_Divergence
Female_Adult	0.707	0.293	0.030
Female_Young	0.887	0.113	0.000
Male_Adult	0.682	0.318	0.157
Male_Young	0.831	0.169	0.006